The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Advances in Low Resource ASR: A Deep Learning Perspective

Hardik B. Sailor, Ankur T. Patil, Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology
(DA-IICT), Gandhinagar-382007, Gujarat, India

{sailor_hardik, ankur_patil, hemant_patil}@daiict.an.in

Abstract

Recently, developing Automatic Speech Recognition (ASR) systems for Low Resource (LR) languages is an active research area. The research in ASR is significantly advanced using deep learning approaches producing state-of-the-art results compared to the conventional approaches. However, it is still challenging to use such approaches for LR languages since it requires a huge amount of training data. Recently, data augmentation, multilingual and cross-lingual approaches, transfer learning, etc. enable training deep learning architectures. This paper presents an overview of deep learning-based approaches for building ASR for LR languages. Recent projects and events organized to support the development of ASR and related applications in this direction are also discussed. This paper could be a good motivation for the researchers interested to work towards low resource ASR using deep learning techniques. The approaches described here could be useful in other related applications, such as audio search.

Index Terms: Low resource languages, ASR, deep learning.

1. Introduction

Speech is being the most natural way of communication, the research community is interested in Human Language Technologies (HLT) for human-machine interaction. It motivates the development of text-to-speech (TTS) and Automatic Speech Recognition (ASR) systems. One of the primitive requirement, for the development of these systems, is sufficient language resources, namely, stable orthography, linguistic expertise, etc. There are more than 7000 languages/dialects spoken in the world, however, only a small fraction of languages offers the resources required for the development of HLT [1]. A language is considered as a low resource (LR) (or under-resource) either in speech, text, phonetic dictionary, or transcribed data (or more than one of these aspects) [2].

Recently, a special class of machine learning called deep learning (or representation learning) produce state-of-the-art results in many speech processing applications including ASR. The detailed survey on using deep learning for speech processing is presented in [3]. However, to train deep architectures, such as Deep Neural Networks (DNN), a large amount of training data is required. Hence, development of ASR using deep learning for LR languages is restricted due to the scarcity of data (in particular, audio, text or both). Still, use of high-resourced and multilingual approaches have emerged to train DNN, and consider LR language as a target language. Some social and cultural aspects bring additional problems in the context of targeted LR language. These problems may include, the languages with many dialectal variations, code-switching or code-mixing phenomena. In unavailability of any under-resourced language, one can borrow resources and knowledge from similar languages. To mitigate the limitation of developing ASR system for LR languages, we will either require innovative data collection/augmentation methodologies to increase the training data or the models for which information is shared amongst languages.

An excellent detailed survey on ASR for LR languages was presented earlier in [4]. The novelty of our overview on ASR for LR languages is that we have presented recent advancements for LR languages using deep learning techniques during last 5 years. Major components of ASR for LR languages are discussed in Section 2. Various project and events organized to support the rapid development of LR language technologies are discussed in Section 3. Finally, Section 4 provides the summary and conclusions of the proposed overview.

2. ASR for Low Resource Languages

2.1. Data Collection and Augmentation

Data collection is an important task for ASR development in LR languages. The data collection approaches can be distinguished as those collected from the available audio resources and audio data collection process [4]. Audio data from the news reading via TV/radio, parliamentary speeches, and the Internet sources are used to build a corpus in the former case. Another approach for data collection includes the recording of speech from speakers of native languages using prompted text material, which reduces the need of manual transcription of the data [4].

Many times, data collection using above mentioned techniques are also very difficult. In such scenarios, data augmentation techniques are used which artificially increase training data directly in signal-domain or feature-domain. To use deep learning approaches, recently data augmentation approach is very popular in low resource ASR. Audio augmentation techniques include tempo and speed perturbation of the original speech data [5]. Feature-domain augmentation includes vocal tract length perturbation (VTLP) and stochastic feature mapping [6]. Various data augmentation techniques were studied for IARPA Babel program and showed improvement in recognition accuracy [7]. Two-stage data augmentation (audio and feature-level, respectively) was also proposed in LR setting [8].

2.2. Feature Learning

Recently, feature learning (also known as representation learning) using neural networks is found to be efficient compared to the traditional Mel filterbank features. Features from the DNN can be taken from the output layer (known as *tandem* features) [9] or low-dimensional hidden layer (known as *bottleneck* features) [10]. The DNN features were proved to be of highly discriminative, robust against the environmental and speaker variations, and language-independent up to certain extent. In the case of ASR for LR languages, DNN representations allow to build acoustic models with limited amounts of training

10.21437/SLTU.2018-4

data. Such an approach also allow sharing of speech data from multiple languages in a bootstrap approach more efficiently for unnoticed languages.

Recently, many studies showed that features obtained from a DNN that was trained with one or multiple languages can be applied to other target LR languages [11–13]. In [14] and [15], authors shown that the data from multiple languages can be used to extract features for a LR language. Data-driven approaches using DNN requires no prior knowledge about linguistic information of the target LR languages. In [16] and [17], the authors shown that a multilingual DNN can be used to initialize a DNN for acoustic modeling based on IPA phone mapping. Substantial improvement was achieved using above mentioned techniques in ASR performance along with its robustness against transcription errors [17]. Bottleneck features were extracted from multitasking DNN in ASR framework and applied in spoken term detection (QbE-STD) for an LR language [18]. Multilingual data selection is possible by obtaining bottleneck DNN features that represent language groups from the training database [19]. Recently, auditory filterbank learning using Convolutional Restricted Boltzmann Machine (ConvRBM) is used along with other features in very recent low resource challenge 2018 organized by Microsoft Research, Banglore, India [20].

2.3. Acoustic Modeling

2.3.1. Use of High Resource Languages

Lacking transcriptions in LR languages motivates the unsupervised or lightly-supervised acoustic modeling. Unsupervised adaptation approaches reduce cost and time in case a prior information of the target LR language is available, such as the language model, pronunciation dictionary, and the identification of an untranscribed speech data. In all such developments, initial ASR systems are built using high-resourced languages followed by adaptation on target LR language that has small amount of training data. Interestingly, even quite dissimilar languages are found to perform very well using above mentioned approach. Current state-of-the-art ASR systems in high-resourced languages typically use context-dependent hybrid DNN-HMM for acoustic modeling and the similar technique is generally used for LR languages.

It is shown in [21] that using English database in a framework of stacked bottleneck DNN improved the BABEL ASR task. Adaptation of an ASR system for LR language is also possible using high resource language by transferring its phoneme inventory as done in [22]. Other studies that used high resource languages are reported in [23, 24]. Such approaches are also called as self-training approaches. In many LR languages, it is possible to find text and audio, however, transcribed speech required for building ASR is unavailable. An interesting study in [25] shows how to use probabilistic transcription LR languages using three approaches, namely, self-learning, mismatched crowd-sourcing, and electroencephalography (EEG).

2.3.2. Multilingual Approaches

Multilingual approaches are generally outperform monolingual approaches for ASR in LR language [26–28]. Detailed discussion on multilingual speech processing applications is given in [29]. Generally, multilingual acoustic modeling is considered as a language-independent approaches. These language-independent approaches can be classified into two categories: (1) generating universal lexicons from a text and (2) acoustic representations from the audio. In [30], authors proposed to create a common set of basic units of a language, which can be represented universally across all the spoken languages. Such basic unit inventory are built from the speech attributes that can be used to build a set of IPA-based language-universal acoustic models [31] or with learning-based techniques. The work proposed in [32] is generally well suited for DNN-based ASR [33].

With the recent success of DNN and their capability to generalize and learn useful acoustic representations of languages, the research moved towards multilingual representations obtained from the DNN. Transcribed speech data from non-target languages can be used to build multilingual DNN acoustic models [34]. Recently, it is shown that multitask learning (MTL) technique shows a notable improvement in error rates over monolingual, multilingual DNN training, semi-supervised learning, and transfer learning framework [35]. MTL is defined as the use of parameters that are shared set in a model trained to optimize the performance metrics from the multiple tasks [35]. In LR scenarios, the multiple tasks could be softmax layers representing multiple LR languages. Advanced topic in ASR field is building end-to-end DNN models where it does not require dictionary, transcription, HMM forced-aligned labels (required in hybrid DNN-HMM setup), or even phones (where character-based models are built). Recently, end-to-end DNN using Connectionist Temporal Classification (CTC) is employed for ASR in LR languages with application to KWS [36].

2.4. Pronunciation Modeling

Pronunciation modeling includes generation of pronunciation dictionary that acts as a link between language model and subword units in the acoustic model. Pronunciation dictionary can be created by grapheme-based approaches where each word is decomposed into graphemes and it is used as a basic unit of the acoustic model [37], [38], [39], [40], [41], [42]. Graphemes-to-phonemes (G2P) models are specifically useful for LR languages that lacks well-developed pronunciation dictionary (also called as *lexicon*). In many cases, when pronunciation dictionary is available, G2P is utilized to obtain out-of-vocabulary (OOV) words pronunciation that do not present in the lexicon. In the earlier work, pronunciation modeling for LR languages were performed using bootstrapping techniques [43,44] or acoustic learning-based techniques [45,46]. Recently, DNN-based approaches have appeared as the novel state-of-the-art for low resource G2P tasks [47,48].

In web approach, word-pronunciation pairs available on web are used to produce pronunciation dictionaries. Wiktionary contains phonetic notations written in the International Phonetic Alphabet (IPA) [49]. Unfortunately, the majority of world's languages are not available in Wiktionary. Studies in [50], [51], and [52] proposed techniques to identify, remove, and substitute erratic pronunciation in the dictionary from the web resources that gave significant improvements. Active learning is also applied in building ASR for OpenKWS and IARPA BABEL program with the addition of web data for language modeling, which is discussed next [53].

2.5. Language Modeling

Statistical language modeling is performed using *N*-gram technique due to their less computational complexity. However, obtaining an *N*-gram statistics required sufficient amount of diverse text data. Interesting experiments conducted in [54] showed the significance of training text material for ASR in LR languages. It also showed a minimum requirement on the amount of text required to build an ASR system in LR languages. Recurrent Neural Network Language Model (RNNLM) is also very effective in language modeling [55]. Earlier, Feedforward Neural Network Language Model (FFNNLM) and RNNLM is applied for language modeling in LR language scenarios [56, 57]. Recently, word-embedding using Long-Short Term Memory (LSTM) is used for LR language modeling [58]. Bi-directional RNNLM is also applied on BABEL database [59].

Conventional hybrid DNN-HMM system block diagram is shown in Figure 1. Generally, RNNLM is used as a rescoring technique (linear interpolation with N-gram) due to its limitation of providing word history during decoding. The block diagram of MTL-DNN for the multilingual approach is shown in Figure 2. Here, target LR language along with other available language database is used to train MTL-DNN system. The output layer of MTL-DNN consists of separate senones labels corresponding to each language under consideration. Next, we discuss projects and events organized to support research in LR languages, specifically in ASR-domain. The summary of few studies for ASR in LR languages is shown in Table 1.

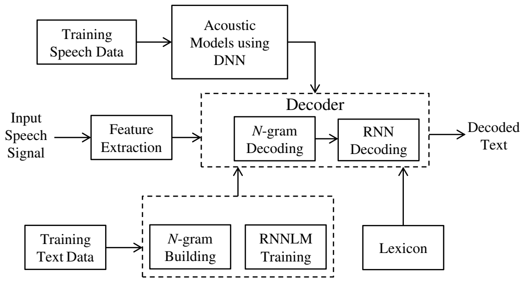

Figure 1: *Block diagram for modern hybrid DNN-HMM system.*

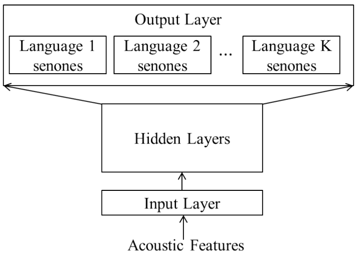

Figure 2: *Block diagram for MTL-DNN system for multilingual ASR. Adapted from [35].*

3. Projects and Events

3.1. IARPA BABEL Program

The aim of the IARPA BABEL program is to build robust and agile ASR technology that can be quickly applied to build any HTL [61]. The BABEL speech database covers a range of diverse languages and is distributed under two categories for each language - the Full Language Pack (FLP) and the Limited Language Pack (LLP). The FLP and LLP include approximately 80 and 10 hours of training speech data, respectively. The speech data is recorded in realistic scenarios, such as conversational telephone speech, over a wide range of acoustic conditions.

3.2. ISCA Special Interest Group: Under-resourced Languages (SIGUL)

SIGUL is a joint special interest group of the European Language Resources Association (ELRA) and of the International Speech Communication Association (ISCA) [62]. This group organizes several ISCA-supported events, namely, Spoken Language Technologies for Under-resourced Languages (SLTU), and Language Resources and Evaluation Conference (LREC).

3.3. Zero Resource Speech Challenges (ZRSC)

The goal of ZRSC is to develop unsupervised discovery of linguistic units from the speech in any unknown language, specially very limited or zero resourced. “Zero resource” refers to a zero linguistic expertise (e.g., orthographic/linguistic transcriptions), not zero information except audio (visual, human feedback, etc.) [63]. Researchers believe that this goal is theoretically reachable since even the 4 years old child spontaneously learn a language without any supervision from language [63].

3.4. Multi-Genre Broadcast (MGB)-3 Evaluation

One of the tracks of the MGB-3 evaluation consists of building an ASR system for a LR target-domain [64]. The evaluations include the use of 1200 hours of Al-Jazeera audio archive available for the initial model construction (earlier used in MGB-2 evaluations [65]). The target language is the Egyptian dialect of Arabic with 5 hours of speech. The research papers published in MGB-3 evaluations were part of ASRU 2017.

3.5. MSR Low Resource ASR Challenge 2018

Recently, matching with the theme of INTERSPEECH 2018 ‘Speech Research for Emerging Markets in Multilingual Societies, Microsoft Research organized a special session and challenge on speech recognition for three low resource Indian languages [66]. The task is to build ASR in one of the three (all or multilingual) LR Indian languages, namely, Gujarati, Tamil, and Telugu. By this challenge, they provided database of three Indian languages for research in LR Indian languages.

3.6. Other Events and Projects

The Defense Advanced Research Projects Agency (DARPA) Low Resource Languages for Emergent Incidents (LORELEI) Program is to advance the state of computational linguistics and HLT to enable rapid and low-cost development of capabilities for LR languages [67]. The project Breaking the Unwritten Language Barrier (BULB) brings together linguists and computer scientists aims at supporting linguists in documenting unwritten languages [68]. The GlobalPhone project provides multilingual database developed in association with the KIT, Germany. The complete database contains (1) audio data, (2) corresponding transcriptions, (3) pronunciation dictionaries, and (4) baseline N-gram language models [69].

4. Summary and Conclusions

This overview paper on recent advances in the development of ASR for LR languages covers major technological progress using deep learning during past 5 years. Various approaches for data processing, acoustic and language modeling, pronunciation dictionary preparation, etc. were discussed. Recent developments shows that multilingual approaches using DNN for acoustic modeling are very promising direction for research in LR languages. It is also observed that RNNLM are very good at estimating language model probabilities either using less text data or through adaptation. This overview also covers projects and events organized to support research in LR language technologies. We hope that the current demand for speech technology will also increase the interest to cover languages with

Table 1: *Summary of some of the studies for LR languages*

Database	Method	% WER Reduction
IARPA BABEL	Data Augmentation for DNN	0.5-1.2 (absolute) [8]
IARPA BABEL	Multitask DNN	4-14.5 (relative) [35]
IARPA BABEL OP3	End-to-End DNN	0.6-1.1 (absolute) [36]
IARPA BABEL	RNN G2P	9-14 (relative) [47]
IARPA BABEL OP3	Active learning DNN	0.2-2.0 (absolute) [53]
MSR Low Resource Challenge 2018	Multilingual DNN	14.93* (absolute) [60]
MSR Low Resource Challenge 2018	AM features, LF-MMI DNN and RNNLM	9.33* (absolute) [20]

*on the Gujarati language

limited or zero-resource using deep learning approaches.

5. Acknowledgments

Authors would like to thank the MeitY, Govt. of India, for two sponsored projects (1) TTS Phase-II and (2) ASR Phase-II.

6. References

[1] 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.

[2] S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap," in *Proceedings of SPECOM*, Moscow, Russia, 2003, pp. 8–15.

[3] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, July 2017.

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," vol. 56. Elsevier, 2014, pp. 85–100.

[5] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, Dresden, Germany, Sept. 2015, pp. 3586–3589.

[6] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, Sept. 2015.

[7] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *INTERSPEECH, Singapore*, 2014, pp. 810–814.

[8] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. M. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *INTERSPEECH*, San Francisco, 2016, pp. 2378–2382.

[9] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, Istanbul, Turkey, 2000, pp. 1635–1638.

[10] F. Grezl, M. Karafiat *et al.*, "Probabilistic and bottleneck features for LVCSR of meetings," in *ICASSP*, April 2007, pp. 757–760.

[11] A. Stolcke, F. Grezl *et al.*, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *ICASSP*, May 2006, pp. 321–324.

[12] L. Tóth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual portability of MLP-based tandem features–a case study for English and Hungarian," in *INTERSPEECH*, Brisbane, Australia, Sept. 2008, pp. 2695–2698.

[13] C. Plahl, R. Schlüter, and H. Ney, "Cross-lingual portability of Chinese and English neural network features for French and German LVCSR," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 371–376.

[14] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *ICASSP*, 2012, pp. 4269–4272.

[15] S. Thomas, S. Ganapathy *et al.*, "Data-driven posterior features for low resource speech recognition applications," in *INTERSPEECH*, Oregon, USA, 2012, pp. 1–4.

[16] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottleneck features and its application for under-resourced languages," in *Spoken Languages Technologies for Under-Resourced Languages (SLTU)*, Cape Town, South Africa, 2012, pp. 1–4.

[17] N. T. Vu, W. Breiter, F. Metze, and T. Schultz, "Initialization schemes for multilayer perceptron training and their impact on ASR performance using multilingual data," in *INTERSPEECH, Oregon, USA*, 2012, pp. 2586–2589.

[18] H. Chen, C. C. Leung, L. Xie, B. Ma, and H. Li, "Multitask feature learning for low-resource query-by-example spoken term detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1329–1339, Dec. 2017.

[19] E. Chuangsuwanich, Y. Zhang, and J. Glass, "Multilingual data selection for training stacked bottleneck features," in *ICASSP*, Shanghai, China, March 2016, pp. 5410–5414.

[20] H. B. Sailor, M. V. S. Krishna, D. Chhabra, A. T. Patil, M. R. Kamble, and H. A. Patil, "DA-IICT/IIITV system for low resource speech recognition challenge 2018," in *INTERSPEECH, Hyderabad*, Sept. 2018, pp. 1–5.

[21] F. Grzl and M. Karafit, "Boosting performance on low-resource languages by standard corpora: An analysis," in *IEEE Spoken Language Technology Workshop (SLT), Athens, Greece*, Dec. 2016, pp. 629–636.

[22] O. Scharenborg, F. Ciannella *et al.*, "Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system: Preliminary results," in *International Conference on Natural Language and Speech Processing (ICNLSSP), Casablanca, Morocco*, 2017, pp. 1–5.

[23] Ö. Cetin, M. Plauché, and U. Nallasamy, "Unsupervised adaptive speech technology for limited resource languages: A case study for Tamil," in *SLTU*, Hanoi, Vietnam, May 2008, pp. 1–5.

[24] J. Lööf, C. Gollan, and H. Ney, "Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system," in *INTERSPEECH*, Brighton, United Kingdom, Sept. 2009, pp. 88–91.

[25] M. A. Hasegawa-Johnson, P. Jyothi *et al.*, "ASR for under-resourced languages from probabilistic transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 50–63, Jan. 2017.

[26] J. Cui, B. Kingsbury, B. Ramabhadran *et al.*, "Multilingual representations for low resource speech recognition and keyword search," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Arizona, USA, Dec. 2015, pp. 259–266.

[27] S. Thomas, M. L. Seltzer *et al.*, "Deep neural network features and semi-supervised training for low resource speech recognition," in *ICASSP*, Vancouver, Canada, May 2013, pp. 6704–6708.

[28] F. Grézl and M. Karafiát, "Combination of multilingual and semi-supervised training for under-resourced languages," in *INTERSPEECH*, Singapore, Sept. 2014, pp. 820–824.

[29] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Elsevier, 2006.

[30] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[31] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *ICASSP*, Hong Kong, China, April 2003, pp. 144–147.

[32] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Special issue on Paralinguistics in Naturalistic Speech and Language, Computer Speech Language*, vol. 27, no. 1, pp. 209 – 227, 2013.

[33] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *ICASSP*, Kyoto, Japan, March 2012, pp. 4169–4172.

[34] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*, Vancouver, BC, Canada, May 2013, pp. 7319–7323.

[35] V. H. Do, N. F. Chen *et al.*, "Multitask learning for phone recognition of underresourced languages using mismatched transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 501–514, March 2018.

[36] A. Rosenberg, K. Audhkhasi *et al.*, "End-to-end speech recognition and keyword search on low-resource languages," in *ICASSP*, New Orleans, LA, USA, 2017, pp. 5280–5284.

[37] P. Charoenpornsawat, S. Hewavitharana, and T. Schultz, "Thai grapheme-based speech recognition," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, 2006, pp. 17–20.

[38] S. Stüker, "Integrating Thai grapheme based acoustic models into the ml-mix framework-for language independent and cross-language ASR," in *SLTU*, Hanoi, Vietnam, May 2008, pp. 1–5.

[39] S. Gizaw, "Multiple pronunciation model for Amharic speech recognition system," in *SLTU*, Hanoi, Vietnam, May 2008.

[40] V.-B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: Application to vietnamese language," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1471–1482, 2009.

[41] M. Killer, S. Stuker, and T. Schultz, "Grapheme based speech recognition," in *EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 3141–3144.

[42] S. Kanthak and H. Ney, "Multilingual acoustic modeling using graphemes," in *EUROSPEECH*, Geneva, Switzerland, Sept. 2003, p. 11451148.

[43] S. Maskey, A. Black, and L. Tomokiya, "Boostrapping phonetic lexicons for new languages," in *International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, Oct. 2004, pp. 1–4.

[44] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *INTERSPEECH*, Brighton, United Kingdom, Sept. 2009, pp. 2851–2854.

[45] G. Chen, D. Povey, and S. Khudanpur, "Acoustic data-driven pronunciation lexicon generation for logographic languages," in *ICASSP*, Shanghai, China, March 2016, pp. 5350–5354.

[46] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework," *Speech Communication*, vol. 80, pp. 1 – 21, 2016.

[47] P. Jyothi and M. Hasegawa-Johnson, "Low-resource grapheme-to-phoneme conversion using recurrent neural networks," in *ICASSP*, New Orleans, LA, USA, March 2017, pp. 5030–5034.

[48] B. Peters, J. Dehdari, and J. van Genabith, "Massively multilingual neural grapheme-to-phoneme conversion," in *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, Copenhagen, Denmark, 2017, pp. 19–26.

[49] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[50] T. Schlippe, S. Ochs, and T. Schultz, "Wiktionary as a source for automatic pronunciation extraction," in *INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 2290–2293.

[51] A. Ghoshal, M. Jansche *et al.*, "Web-derived pronunciations," in *ICASSP*, Taipei, Taiwan, 2009, pp. 4289–4292.

[52] T. Schlippe, S. Ochs, N. T. Vu, and T. Schultz, "Automatic error recovery for pronunciation dictionaries," in *INTERSPEECH*, Sept. 2012, pp. 2298–2301.

[53] A. R. Syed, A. Rosenberg, and M. Mandel, "Active learning for low-resource speech recognition: Impact of selection size and language modeling data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 5315–5319.

[54] T. Pellegrini and L. Lamel, "Investigating automatic decomposition for ASR in less represented languages," in *INTERSPEECH-ICSLP*, Pittsburgh, Pennsylvania, Sept. 2006, pp. 1–4.

[55] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH, Makuhari, Chiba, Japan*, 2010, pp. 1045–1048.

[56] A. Gandhe, F. Metze, and I. Lane, "Neural network language models for low resource languages," in *INTERSPEECH*, Singapore, Sept. 2014, pp. 2615–2619.

[57] A. Ragni, E. Dakin, X. Chen, M. J. Gales, and K. M. Knill, "Multi-language neural network language models," in *INTERSPEECH*, San Francisco, Sept. 2016, pp. 3042–3046.

[58] O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, "Cross-lingual word embeddings for low-resource language modeling," in *Proceedings of the 15^{th} Conference of the European Chapter of the Association for Computational Linguistics (ACL)*, vol. 1, Valencia, Spain, 2017, pp. 937–947.

[59] X. Chen, A. Ragni, X. Liu, and M. J. Gales, "Investigating bidirectional recurrent neural network language models for speech recognition," in *INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 269–273.

[60] B. Pulugundla, M. K. Baskar *et al.*, "BUT system for low resource Indian language ASR," in *INTERSPEECH, Hyderabad*, Sept. 2018, pp. 1–5.

[61] IARPA, "The IARPA BABEL program," URL: https://www.iarpa.gov/index.php/research-programs/babel, {Last Accessed: 25 July 2018}.

[62] ISCA and ELRA, "ISCA Special Interest Group: Under-resourced Languages (SIGUL)," URL: https://www.isca-speech.org/iscaweb/index.php/sigs?layout=editid=198, {Last Accessed: 25 July 2018}.

[63] Organizers of ZeroSpeech challenges, "The zero resource speech challenge," URL: http://sapience.dec.ens.fr/bootphon/index.html, {Last Accessed: 25 July 2018}.

[64] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *ASRU*, Okinawa, Japan, Dec. 2017.

[65] A. Ali, P. Bell *et al.*, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," in *SLT*, San Diego, California, Dec. 2016, pp. 279–284.

[66] Microsoft, "INTERSPEECH 2018 special session: Low resource speech recognition challenge for Indian languages," URL: https://www.microsoft.com/en-us/research/event/interspeech-2018-special-session-low-resource-speech-recognition-challenge-indian-languages, 2018, {Last Accessed: 25 July 2018}.

[67] DARPA, "Low resource languages for emergent incidents (LORELEI)," URL: https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents, {Last Accessed: 25 July 2018}.

[68] G. Adda, S. Stker *et al.*, "Breaking the unwritten language barrier: The BULB project," in *SLTU*, Yogyakarta, Indonesia, 2016, pp. 8 – 14.

[69] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A multilingual text and speech database in 20 languages," in *ICASSP*, May 2013, pp. 8126–8130.