M2ASR: AMBITIONS AND FIRST YEAR PROGRESS

Dong Wang[1], *Thomas Fang Zheng*[1], *Zhiyuan Tang*[1], *Ying Shi*[1], *Lantian Li*[1], *Shiyue Zhang*[1]
Hongzhi Yu[2], *Guanyu Li*[2], *Shipeng Xu*[2]
Askar Hamdulla[3], *Mijit Ablimit*[3], *Gulnigar Mahmut*[3]

[1] CSLT, RIIT, Tsinghua University
[2] Northwest Minzu University
[3] Xinjiang University

ABSTRACT

In spite of the rapid development of speech techniques, most of the present achievements are for a few major languages, e.g., English and Chinese. Unfortunately, most of the languages in the world are 'minority languages', in the sense that they are spoken by a small population and with limited resource accumulation. The present speech technologies are mostly based on deep learning and big data, therefore not directly applicable to minority languages. However, minority languages are so numerous and important that if we want to break the language barrier, they must be seriously taken into account.

Recently, the Chinese government approved a fundamental research for minority languages in China: Multilingual Minorlingual Automatic Speech Recognition (M2ASR). Although the initial goal was speech recognition, the ambition of this project is more than that: it intends to publish all the achievements and make them free for the research community, including speech and text corpora, phone sets, lexicons, tools, recipes and prototype systems. In this paper, we will describe this project, report the first-year progress, and present the future plan.

Index Terms— Minority language, speech recognition, data resource, Uyghur, Kazak, Mongolia, Kirgiz, Tibetan

1. INTRODUCTION

Speech technologies, including speech recognition (ASR), speaker recognition (SID), and text to speech synthesis (TTS) among others, have achieved brilliant progress in recent years, mostly attributed to the development of the deep learning approach [1]. A key ingredient of the success of deep learning is a large amount of training data, so that the model can learn rules from raw signals with little prior knowledge. For that reason, the most impressive progress of speech techniques took place with 'major languages', i.e., languages spoken by large populations and whose data and other resources are well archived. Typical major languages include English and Chinese. However, for most languages in the world, the resources are very limited, not only in speech and text data, but also in the basic acoustic and linguistic research, e.g., phone sets, lexicons, morphological rules. These languages are called 'minority languages'. Due to the lack of resources, applying modern speech technologies to minority languages remains a challenging task. Ironically, it is the people in minority nations that mostly require speech technologies to break the language barrier. Moreover, if we want to protect the diversity of human culture, the best approach to the protection is to allow people communicating with others using their own native languages. All the above highlight the importance of speech processing research for minority languages.

For the situation of China, the Han people is the largest population (1.2 Billion in 2010[1]) and most of them speak Mandarin Chinese. Besides, there are 55 minority languages, and most of them have their own languages. There are about 72 languages actively used, covering 5 language families. For most of the languages, the resources are highly limited. Besides a very few exceptions (e.g., Uyghur, Tibetan), the speech and text databases are pretty scarce and far from standard, and the acoustic and linguistic research is also far from extensive. On the other hand, almost all the minority languages are significantly influenced by Chinese, and languages in the same region impact each other, e.g., Uyghur and Kazak. This leads to special acoustic/linguistic variety, particularly for languages without their own writing systems. The intermingling of data sparsity and acoustic/linguistic variety causes significant difficulty in speech research for minority languages, impeding the migration of modern speech technologies.

To boost the speech technologies for minority languages, the China government recently approved a 5-year project through its NSFC (National Natural Science Foundation of China) program. This project, called Multilingual Minorlingual Automatic Speech Recognition (M2ASR), aims to construct a multilingual speech recognition system for five minority languages in China (Tibetan, Mongolia, Uyghur, Kazak and Kirgiz). Three participants are involved in the

[1] https://en.wikipedia.org/wiki/List_of_ethnic_groups_in_China_and_Taiwan

project: Tsinghua University, Northwest Minzu University, and Xinjiang University. Although the original goal was speech recognition, our ambition is beyond that scope: we hope to construct a full set of resources for the 5 languages, including speech and text databases, lexicons, acoustic and linguistic rules, recipes and prototype systems. By making all the resources open and free, we hope to benefit researchers in broader areas related to minority language research. More information can be found in the project web site (http://m2asr.cslt.org).

2. M2ASR PROJECT

In general, the goal of the M2ASR project is two-fold: firstly, it intends to develop a multilingual speech recognition system that can conduct ASR for the five target languages simultaneously; secondly, it will publish a set of resources for the five target languages to assist research on related areas.

2.1. Multilingual speech recognition

Deep neural networks (DNNs) have an inherent capability for multilingual speech recognition [2, 3, 4]. The insight is that human languages share some commonality in both acoustic and phonetic aspects, and so patterns at some levels of abstraction can be shared. Inspired by this insight, it is possible to train a multilingual DNN whose low-level layers are shared while the high-level layers are language specific [2]. By this architecture, the low-level layers can learn shared features, while the high-level layers will learn language-dependent phone classifiers. This feature-sharing approach is particularly useful for minority language ASR, as it allows reusing a feature extractor trained with data in a rich-resource language.

Despite the brilliant success of the feature-sharing approach, it is still far from sufficient to solve the multilingual ASR problem. We highlight three issues that we will try to address with the M2ASR project.

2.1.1. Insufficient sharing

The popular feature-sharing approach works at the acoustic layer, which is plausible but not sufficient. The shortage of resources of minority languages is not only in speech data, but also in lexicon and text data. It is well known that languages spoken by people located in proximal regions impact each other, and share words and concepts. This provides a possibility to use a shared word/concept space, based on which language models can be built for each individual language. This is illustrated in Fig. 1, where the words of a particular language are categorized into two types: entity words and functional words. The entity words represent concepts, while the functional words glue entity words to form a legal sentence. The functional words and their compositional rules are language specific, while the entity words can be shared across languages. This results in a concept space that can be used to improve the language modeling. The word sharing can be established by either a language-to-language dictionary, or the cosine similarity of word embeddings [5, 6]. If it is based on word embedding, then a neural language is perhaps more appropriate [7].

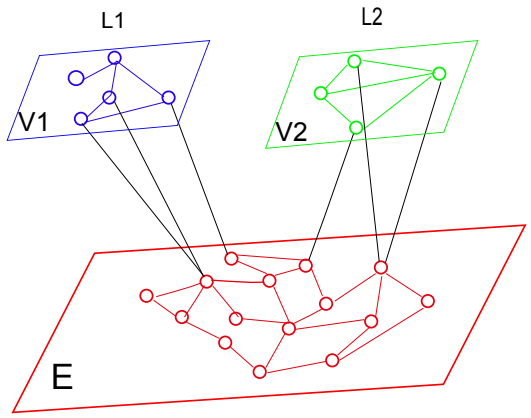

Fig. 1. The LM sharing approach. The entity words form a shared concept space E, and the functional words form language-specific space V.

2.1.2. Language independent decoding

The multilingual ASR based on feature-sharing is in fact not true multilingual, as in the recognition (decoding) phase, the multilingual DNN is still used independently to conduct monolingual recognition, using the language model of that language. A true multilingual ASR should deal with multilingual input, without knowing what the language is in prior. A possible approach is to mix the language models of multiple languages and resort to the decoding strategy (e.g., beam search) to select the best language. However, this will decrease the performance significantly, due to the inter-language path competition within the limited beam width. Another approach is to determine the language by a language identification (LID) system, and then invoke the correct monolingual decoding (using the multilingual DNN). This, however, will lead to unacceptable latency. The third approach is to invoke multiple monolingual decoding processes, each for a particular language, and then select the results by either LID or confidence ranking. This approach demands more computation.

A possible solution is to decode in the mixlingual decoding space (by mixing monolingual LMs), but regularize the decoding by a frame-level language prior. This is illustrated in Fig. 2. Our preliminary study demonstrated that this language-aware multilingual decoding is possible [8], where the frame-level language prior is produced by a recurrent neural network (RNN).

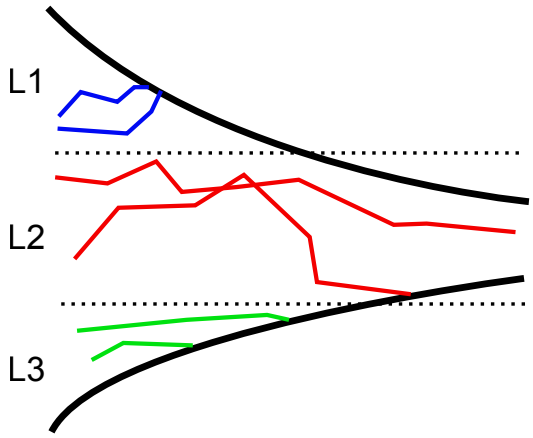

Fig. 2. Multilingual speech recognition with frame-level language information. At the beginning, the priors for the languages are identical, and paths in various languages are equally distributed in the search beam (the area between the two black curves). With the decoding continuing, the correct language obtains a larger prior (obtained from a frame-level LID system), which strengthens the paths in that language (red curves) and attenuates the paths in incorrect languages (blue and green curves).

2.1.3. Difficulty with very low-resource languages

Another problem for minority language ASR is that the resources of some languages are very rare: there might be no standard phone sets, no reasonable lexicons, and even no writing systems. For these languages, the traditional ASR architecture based on phone discrimination and phone to word conversion does not work. A possible solution is the end-to-end training, with which the input end is the speech signal in the minority language, and the output end is the word sequence of a major language, e.g., Chinese. This approach essentially integrates speech recognition and machine translation (MT) and trains them in a joint way. A potential problem for this end-to-end system is that the alignment between the speech frames in the minority language and the phone/word sequences in the major language might be difficult or even impossible. A potential solution is the attention-based mechanism that was proposed by [9] and has been utilized in ASR [10, 11]. With the attention mechanism, the alignment can be soft and be learned together with the recognition model.

2.2. Free data program

Data resources are highly important for ASR and other speech processing tasks. For the minority languages considered by M2ASR, the data resources are still limited, in both speech and text. There are only several databases, but almost all of them are held for private usage. There are nearly no standard databases, not to mention free ones. A known exception is the THUYG20 speech corpus published by Tsinghua University. This 20-hour Uyghur speech database [12, 13] can be downloaded freely from OpenSLR[2], and the associated Kaldi recipe was also published on GitHub[3].

Considering the predominant importance of free data, the second goal of the M2ASR project was set to publish a full set of speech and text databases for all the covered languages. For the speech data, we will publish a seed set and a body set for each language. The seed set consists of 50 hours of clean speech in reading style, spoken by 100 persons. The body set consists of 200 hours of spontaneous speech, spoken by 300 persons. For the text data, we will publish a text collection of 50M words for each language. All the data will be free.

Besides the data, M2ASR will also publish the acoustic and linguistic resources, particularly phone sets and lexicons. We will also publish a set of tools for minority language processing, including text normalization, morphology analysis, code conversion. Kaldi recipes and prototype systems will be also published.

3. FIRST YEAR PROGRESS

The M2ASR was started in September, 2016. During the one-year period, a multitude of achievements have been attained. This section will report our current status, what has been achieved and what is on going.

3.1. Research

A multitude of research has been conducted following the goal of M2ASR, on both the acoustic part and the linguistic part. We will report two representative studies. More information can be found in the publication page of the project.

3.1.1. Phonetic temporal neural (PTN) model for LID

A key research we conducted in the past year is a high-quality LID system. To achieve the language-aware multilingual decoding, a short-segment LID is necessary, frame-level the best. Conventional methods are based on probabilistic models, either on linguistic units [14] or acoustic features [15, 16]. These methods suffer from significant latency. Recently developed DNN/RNN approach works well for frame-level discrimination [17, 18], but the accuracy is still unsatisfactory and the performance is vulnerable to noise and channel variation. We hypothesize that a major problem of the existing neural approach is that the discriminative capability of these 'LID oriented DNNs' is limited. The output only consists of a few target language, which can not provide sufficient information to train a robust model. Moreover, the temporal property among acoustic units is nearly ignored, though this knowledge was successfully used in the historical

[2]http://www.openslr.org/22/

[3]https://github.com/wangdong99/kaldi

probabilistic approach, e.g., the famous Phone Recognition Language Modeling (PRLM) framework [14].

We proposed a phonetic temporal neural (PTN) model, which involves a phonetic DNN to produce phonetic features and an LID RNN that accepts the phonetic input and generates frame-level language posteriors. By this architecture, phone labels rather than language labels are used to train the phonetic DNN, which provides more discriminative information for the model. Although the model (phonetic DNN) does not target for the goal task (language discrimination), it still provides rich information. With this information, the LID RNN can learn the temporal property as the language model in the PRLM framework. Our experiments demonstrated that this PTN approach works fairly well [19].

3.1.2. Memory structure for OOV treatment

Another major problem for minority language ASR is the high rate of out-of-vocabulary (OOV) words. This high OOV rate is partly attributed to the lack of lexicon resource, but it is also related to the nature of the language. For M2ASR, Uyghur, Kazak, and Kirgiz are all agglutinative languages, for which a word consists of a stem and unlimited suffixes, leading to a nearly unlimited lexicon. This poses a big challenge for language modeling, as no matter how large the lexicon is, the OOV rate is still significant. Moreover, even if the words are in the lexicon, most of them are rare words and cannot be well represented by the LM.

We proposed a memory-based approach to attack the OOV problem [20]. The key idea is to establish a mapping table that maps OOV or rare words to more frequent words, so that a back-up LM can be trained with stronger statistics. The back-up LM and the original LM are then combined to conduct decoding. This approach has been tested on a Uyghur-Chinese translation task and showed significant performance improvement [20]. It will be straightforward to extend the architecture to ASR. In fact, the central idea that using similar words to represent OOV and rare words have been developed in our previous work on ASR [21], although in that paper we did not consider the large proportion of OOVs caused by the agglutinative nature.

3.2. Data release

Following the data free program, we release the resources shown in Table 1 as the first-year output. Some of the resources can be downloaded from the project web site directly, and some need email request.

3.3. Toolkit

We also published a toolkit called 'M2LP' that can be used to conduct text processing for the five languages covered by M2ASR [24]. It was initially designed as a general framework that assists morphological analysis for agglutinative languages. As have been discussed in previous sections, Altaic languages, including Uyghur, Kazk, and Kirgiz involve rich morphological structures. If we ignore these structures and follow the regular LM recipe to build word-based LMs, then rare and OOV words will be a significant problem. A possible solution is to split words into morphemes and construct morpheme LMs, as has been done in our previous work on Uyghur ASR [13]. However, this morphological analysis is not easy, due to the complex variation in both spelling and pronunciation [25]. M2LP provides a framework that decomposes the complexity into language-dependent rules and a language-independent inference process.

Besides the morphological analysis, M2LP also implemented/will implement modules for code change, spell normalization, clustering, and similar word substitution. This toolkit is free and can be download from GitHub[4].

Language	Resource	Amount	Ref.	Avail.
Uyghur	Speech	50 hours	[12]	Req.
	Lexicon	45k words		Pub.
	Recipe & Proto			Pub.
Kazak	Speech	50 hours	[22]	Req.
	Lexicon	100k words		Pub.
	Recipe & Proto			Pub.
Mongolian	Transcription	60k sentences	[23]	Pub.
	Lexicon	26k words		Pub.
Tibetan	Speech	20 hours	[22]	Req.
	Text	310k sentences		Pub.
	Lexicon	6013 syllables		Pub.
	Recipe & Proto			Pub.

Table 1. M2ASR First-year resource release. 'Req.' means the resource can be obtained by email request, 'Pub.' means it can be downloaded freely from the project web site.

4. CONCLUSIONS

We have introduced the NSFC key project M2ASR. Although the initial goal of this project was a multilingual minorlingual ASR system, we hope it can produce more for the research community. Firstly, we will publish all the speech and text databases for free for research purpose, so that researchers have sufficient and standard data to conduct their research; Secondly, we will publish the standard baseline results and the recipe that can reproduce these results, so that researchers have a benchmark system to evaluate their research; thirdly, we will publish all the data, tools, systems, progress, documents, and publications on m2asr.cslt.org, so that new researchers can follow us to perform the research from scratch. The last but not the least, our publications will benefit not only the ASR research, but also related areas including speech

[4]https://github.com/M2ASR

synthesis, acoustics and linguistics, and even machine translation.

This project is 5-years long. In the next year, we will complete the publication of the seed (reading style) speech databases for all the five languages, and start to construct spontaneous speech databases. The text databases will be also finished and published. The Mogolia and Kirgiz recipes will be ready for publication, and the initial bi-lingual ASR system should be done.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under the project No.61371136, No.61633013 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

5. REFERENCES

[1] Li Deng and Dong Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[2] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013, pp. 7304–7308, IEEE.

[3] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013, pp. 7319–7323, IEEE.

[4] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, M Ranzato, Matthieu Devin, and Jeffrey Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2013, pp. 8619–8623, IEEE.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[7] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*, pp. 137–186. Springer, 2006.

[8] Zhiyuan Tang, Lantian Li, and Dong Wang, "Multi-task recurrent model for speech and speaker recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–4.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[10] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[11] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.

[12] Aisikaer Rouzi, YIN Shi, ZHANG Zhiyong, WANG Dong, Askar Hamdulla, and ZHENG Fang, "Thuyg-20: A free uyghur speech database," *Journal of Tsinghua University (Science and Technology)*, vol. 57, no. 2, pp. 182–187, 2017.

[13] Askar Rozi, Dong Wang, Zhiyong Zhang, and Thomas Fang Zheng, "An open/free database and benchmark for Uyghur speaker recognition," in *Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015 International Conference*. IEEE, 2015, pp. 81–85.

[14] Marc A Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, pp. 31, 1996.

[15] Marc A Zissman, "Automatic language identification using Gaussian mixture and hidden Markov models," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1993, vol. 2, pp. 399–402.

[16] Najim Dehak, A.Torres-Carrasquillo Pedro, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proceedings of the Annual Conference of International Speech*

Communication Association (INTERSPEECH), 2011, pp. 857–860.

[17] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno, "Automatic language identification using deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 5337–5341, IEEE.

[18] Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Hasim Sak, Joaquin Gonzalez-Rodriguez, and Pedro J Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014, pp. 2155–2159.

[19] Zhiyuan Tang, Dong Wang, Yixiang Chen, Lantian Li, and Andrew Abel, "Phonetic temporal neural model for language identification," *arXiv preprint arXiv:1705.03151*, 2017.

[20] Shiyue Zhang, Gulnigar Mahmut, Dong Wang, and Askar Hamdulla, "Memory-augmented Chinese-Uyghur neural machine translation," *arXiv preprint arXiv:1706.08683*, 2017.

[21] Xi Ma, Dong Wang, Javier Tejedor, Xi Ma, Dong Wang, Javier Tejedor, Javier Tejedor, Xi Ma, and Dong Wang, "Similar word model for unfrequent word enhancement in speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1819–1830, 2016.

[22] Guanyu Li, Hongzhi Yu, Thomas Fang Zheng, and Jinghao Yan, "Free linguistic and speech resources for tibetan," in *APSIPA 2017 submitted*, 2017.

[23] Shipeng Xu, Hongzhi Yu, Thomas Fang Zheng, and Jinghao Yan, "Language resource construction for mongolian," in *APSIPA 2017 submitted*, 2017.

[24] Mijit Ablimit, Sardar Parhat, Askar Hamdulla, and Thomas Fang Zheng, "A multilingual language processing tool for Uyghur, Kazak and Kirghiz," in *APSIPA 2017 submitted*, 2017.

[25] Mijit Ablimit, Tatsuya Kawahara, and Askar Hamdulla, "Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language," *Speech communication*, vol. 60, pp. 78–87, 2014.