Nict-Tib1: A Public Speech Corpus Of Lhasa Dialect For Benchmarking Tibetan Language Speech Recognition Systems

Conference Paper · November 2022

DOI: 10.1109/O-COCOSDA202257103.2022.9997917

3 authors:

Kak Soky
Kyoto University
11 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Zhuo Gong
The University of Tokyo
8 PUBLICATIONS **11** CITATIONS
SEE PROFILE

Sheng Li
Institute of Science Tokyo
138 PUBLICATIONS **999** CITATIONS
SEE PROFILE

O-COCOSDA 2022, Hanoi, Nov 24-26, 2022

NICT-TIB1: A PUBLIC SPEECH CORPUS OF LHASA DIALECT FOR BENCHMARKING TIBETAN LANGUAGE SPEECH RECOGNITION SYSTEMS

Kak Soky[⋆‡], *Zhuo Gong*[†‡], *Sheng Li*[‡]

[⋆]Kyoto University, Kyoto, Japan
[†]The University of Tokyo, Tokyo, Japan
[‡]National Institute of Information and Communications Technology (NICT), Kyoto, Japan

ABSTRACT

The Lhasa dialect is the primary Tibetan dialect, with the most speakers in Tibet and the most extensive written scripts over its lengthy history. Studying speech recognition methods in the Lhasa dialect significantly conserves Tibet's distinctive linguistic variety. Previous research on Tibetan speech recognition focused on academic research on non-public datasets, e.g., selecting phone-level acoustic modeling units and incorporating tonal information, but had less contribution to limited data for the community. To solve the low-resource data problem, we introduce the NICT-Tib1 (phase1) database, a new open-sourced database for the Lhasa dialect. We further update benchmark systems under the monolingual and multilingual settings, respectively. Experimental results show that the performances of these models are consistent with previous work. We believe our work will promote the existing speech recognition research on the Tibetan language, and other low-resource languages.

Index Terms— Speech recognition, Tibetan language, Lhasa dialect, low-resource data

1. INTRODUCTION

Tibet's culture is going through a significant modernizing shift in the twenty-first century. Today, one of the most challenging issues is how to protect Tibet's distinctive linguistic uniqueness. There are three main varieties of Tibetan spoken today: Lhasa Tibetan, Khams Tibetan, and Amdo Tibetan. The most popular dialect and one with the most significant percentage of speakers (90%) are the Lhasa (central Tibetan dialect). This language has been used for a long time to write the majority of historic Tibetan texts. For this reason, studying how to apply automatic speech recognition (ASR) techniques to the Lhasa dialect has special meanings. Our previous works have been done consistent and systematic research on applying speech recognition technology to the Tibetan language.

Early work on conventional phone-based [1, 2] Tibetan ASR speech recognition research focused on selecting acoustic modeling units [3], incorporating effective tonal information [4]. When investigating End-to-End ASR models for Tibetan language [5, 6, 7, 8, 9, 10], we introduced highly compressed, and reliable sub-character units for acoustic modeling which have never been used before [11]. Later, the transfer-learning[12] and meta-learning[13] are further used to improve the ASR performances. In multilingual settings, we got second place in the world for Tibetan ASR performance in OLR2021 (oriental language recognition challenge), in which the overall ASR performance of 13 languages is third place [14]. The multilingual (including Tibetan) ASR model is used to enhance the language identification system [15]. The above End-to-End modeling technologies also proved effective in ASR tasks with radicals of Chinese characters [16], articulatory units [17], Southeastern Asian languages [18] and disordered speech [19].

In this paper, we introduce the NICT-Tib1 (phase1) database[1], a new open-sourced database for the Lhasa dialect, to further solve the low-resource data problem and contribute to the community. We build concrete ASR systems with the Transformer in monolingual and multilingual settings.

The remainder of this paper is structured as follows. The related works are overviewed in Section 2. Section 3 explains and evaluates benchmark systems with our dataset. This paper concludes in Section 4.

2. RELATED WORKS

2.1. Background Knowledge of Tibetan Language

2.1.1. Historical Influence

1. It was one of the five official languages of the Qing Empire (i.e., Mandarin, Manchurian, Mongolian, Tibetan, Uyghur) for over two hundred years. Almost all of the empire's official documents had five parallel recordings.

This work was done during Kak Soky and Zhuo Gong's internship in NICT.

[1]We are also holding another larger Tibetan database under processing, namely NICT-Tib2 (phase2), which will soon be released.

2. It belongs to the Sino-Tibetan language family and shares similarities with other languages in this family (e.g., Chinese and Myanmar), as shown in Fig.1. It preserves the ancient form of the pronunciation of the Sino-Tibetan language. Thus, it has special meaning in uncovering the evolution of the Chinese language.

$$\begin{cases} Indo-European \mapsto \begin{cases} Germantic \mapsto English \\ Italic \mapsto Spanish, French, Italian \end{cases} \\ Sino-Tibetan \mapsto \textbf{Mandarin}, Myanmar, \textbf{Tibetan} \\ Altaic \mapsto Japanese, Korean \\ Austric \mapsto Vietnamese, Thai, Indonesian \end{cases}$$

Fig. 1: A brief summary of language family tree.

3. Since Mongolian and Tibetan people shared the same Buddha religion, massive Mongolian religious scripts were written in Tibetan. Moreover, the Tibetan alphabet can precisely spell most Indian languages (e.g., Devanagari). It also inspired the invention of the Korean alphabet.

2.1.2. Tibetan Accents and Geological Distributions

As mentioned above, there are three dialects of the Tibetan language. They are Lhasa, Khams, and Amdo, and the geological distributions are shown in Figure 2.

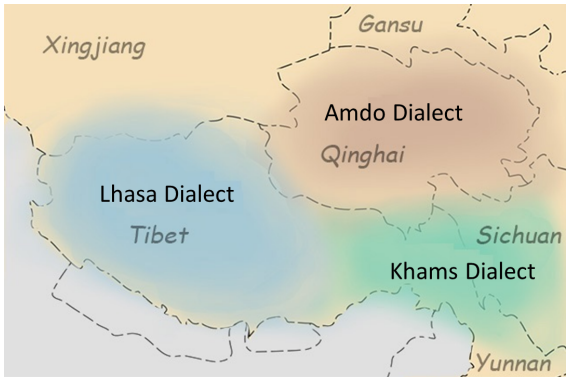

Fig. 2: Geological distribution of major Tibetan dialects.

2.1.3. Tibetan Writing System

As shown in Figure 3, a typical Lhasa Tibetan character is composed of root-script (Root.), pre-script (Pre.), super-script (Super.), sub-script (Sub.), vowels (Vo.), and post-scripts (Post.). Their combination results in enormous vocabulary. The order of pronunciation is Pre. → Super. → Root. → Sub. → Vo. → Post.

As shown in Figure 3, speech recognition performance can be affected by how these elements are combined to define the phone set. Additionally, the pronunciation of these components is where the initials originate from. The actual initials are 28, according to the Lhasa dialect. The potential pairings of a character's vowel and its post-scripts determine Tibetan finals.

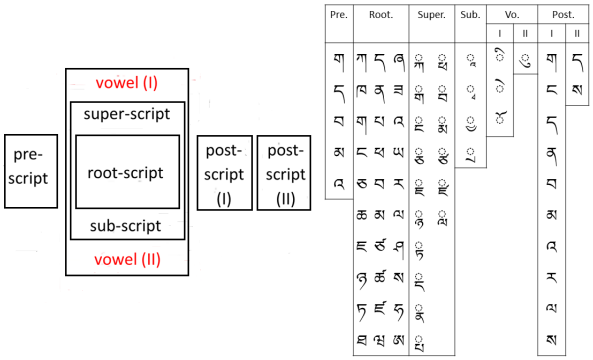

Fig. 3: The root-script (Root.), pre-script (Pre.), super-script (Super.), sub-script (Sub.), vowels (Vo.), and post-scripts (Post.) make up the construction of a letter in the Lhasa Tibetan writing systems. The order of pronunciation is Pre. → Super. → Root. → Sub. → Vo. → Post.

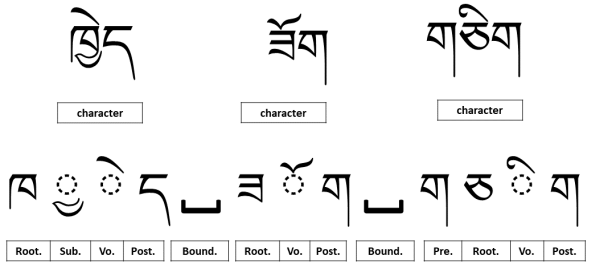

Fig. 4: Segmenting characters to sub-character units.

As shown in Figure 4, Tibetan characters can be further segmented into sub-characters, which can be used as End-to-End modeling units.

2.2. Our Previous Lhasa-Tibetan ASR systems

GMM/DNN-HMM hybrid system: The traditional GMM-HMM[1] and DNN-HMM[2] automatic speech recognition (ASR) systems are phone-based and need individually adjusted acoustic models, lexicon, and language models. The initial/final based non-tonal phone set was selected as an acoustic modeling unit in an earlier study [4, 3]. The non-tonal phone set was constructed using the results of earlier phonological research of the Lhasa spoken language[20]. There are 48 final units and 29 initial consonants without considering the tones. Since Lhasa Tibetan does not yet have a definitive tonal pattern, a four-tone pattern based on the four

contour contrasts scheme [4] was developed. The 192 tonal finalists are added to the 48 non-tonal finals. The 29 initials remain the same. The filterbank in that work did not take advantage of the pitch-related features.

Due to the issue with the data's low resource availability, the improvement is still relatively modest. To improve Tibetan speech recognition performance, we are required to research innovative acoustic modeling approaches.

End-to-End system: The End-to-End neural network model, which obtained encouraging results on ASR tasks, simplified the development of the ASR system and resolved the sequence labeling problem between variable-length speech frame inputs and label outputs (e.g., phone, character, syllable, word, sub-word). The transformer[10, 21, 22, 23, 24], lattice-free MMI (LFMMI) [25], connectionist temporal classification (CTC) [5, 6], attention-based encoder-decoder (Attention) models [7, 8], jointly trained models for CTC and attention objectives (CTC/Attention) [9, 26, 27, 28], and other recent End-to-End models (e.g., Wav2vec2.0 [29]) have all been investigated.

For Tibetan ASR, the sub-character unit also proved effective for modeling Lhasa-dialect under low-resourced conditions [11]. As we know, a Tibetan character is further segmented into a sequence of sub-character tokens, as shown in Figure 4. Further research shows that we can build concrete End-to-End ASR systems combining different level units using transfer-learning[12] and meta-learning[13].

We further apply the Tibetan language ASR technology in multilingual settings. We got second place in the world for Tibetan ASR performance in OLR2021 (oriental language recognition challenge), in which the overall ASR performance of 13 languages is third place [14]. We also discovered that the multilingual (including Tibetan) ASR model could be used to effectively initialize the language identification system [15].

3. THE LHASA DIALECT TIBETAN END-TO-END ASR SYSTEMS WITH NICT-TIB1

In this study, based on transformer models, we develop Lhasa dialectal End-to-End ASR systems utilizing NICT-Tib1 in monolingual and multilingual settings.

3.1. Monolingual Transformer-based system

Experimental Settings: The speech corpus has 33.5 hours of speech signal data collected from 20 Tibetan Lhasa native speakers, including ten males and ten females. All the speakers are native speakers of the Lhasa dialect living in Tibet. The speech signal is sampled at 16KHz with 16-bit quantization. The recording scripts consist of mainly declarative sentences covering broad topics to build a practical ASR system. The recordings are collected from cell phones in a clean environment. There are 15,646 sentences in the corpus (the extremely long and short sentences are removed). When post-editing the speech data, we hide the speaker's identifiable information contained in the speech, strictly following the GDPR law and increasing privacy concerns [30].

We implemented the model using a Transformer-based architecture of the ESPnet [26]. We employed 3-dim pitch and 80-dim log-Mel filterbank coefficients following the typical configuration. To enhance voice data, we used speech perturbation and SpecAugment. Six encoder layers and six decoder layers make up the network. The dropout was set at 0.1, and the feed-forward network's size was 2048. The model made use of 256-dimensional, four-head self-attention. A two-layer time-axis convolutional layer with 256 channels, a stride size of 2, and a kernel size of 3 were used to begin this network with downsampling. With one V100 GPU and batch size 64, the model was jointly trained with CTC (weight $\alpha = 0.3$) for 30 epochs. The Noam optimizer was utilized with 25000 warm-up steps and 5 as the initial learning rate. The vocabulary size is 85.

Moreover, we trained DNN-HMM baseline models using the same setting of previous work [11, 12].

Table 1: Speech corpus of Lhasa dialect

Datasets	#Speakers	#Utterances	Hours
Training	15	14,438	29.5
+speed perturbation	15	43,314	88.5
Development (Dev)	2	1,150	2.0
Testing (TST)	3	1,058	2.0

Table 2: ASR performance (character error rate (CER%) as the criterion for evaluation) of transformer-based models.

Network	#unit	CER%	
		Dev	TST
DNN-HMM	Senome 3000	/	7.2%
Transformer	Char. 85	6.5%	7.8%
+SpecAugment		6.2%	6.8%

Experimental Results: The model trained with 85 character units achieved the closest performance with the highly optimized baseline DNN-HMM model, which was enhanced with the language model on TST. The small performance gap between the transformer and baseline DNN-HMM system can be compensated with the SpecAugment. This result is consistent and reasonable with previous work.

3.2. Multilingual Transformer-based system

Experimental Settings: We further mix the NICT-Tib1 Tibetan data with part of the training set of OLR (oriental

Table 3: Multilingual OLR datasets (AP-17 clean channel)

Language	tb	zh	ct	vi	ja	ko	id	ru	kz	uy
#Vocabulary	85	2842	3081	142	1788	1139	53	65	87	33
OLR-Train (hours)	8.8	6.8	6.9	7.6	5.1	2.6	6.6	4.6	3.1	8.8
OLR-Test (hours)	1.2	0.8	0.8	0.8	0.7	0.3	0.8	0.4	0.4	1.2

Table 4: ASR results for systems with different settings on OLR-Test (character error rate (CER%) as the criterion for evaluation).

	CER% of OLR-Test									
Language	tb	zh	ct	vi	ja	ko	id	ru	kz	uy
OLR-Train	6.9	46.8	40.2	18.0	38.6	48.4	17.5	27.5	34.1	20.0
OLR-Train + NICT-Tib1	6.3	39.4	33.4	17.1	34.1	43.0	16.7	26.8	32.6	20.6

language recognition challenge) [14]. We selectively use the AP-17 dataset, which has both word labels and language tags for every sentence. The AP-17 has ten languages: Kazak (kz), Tibetan (tb), Uyghur (uy), Cantonese (ct), Indonesian (id), Japanese (ja), Korean (ko), Russian (ru), Vietnamese (vi), Mandarin (zh). Moreover, it includes both clean and noisy channel data, and we only use the clean channel data to match with NICT-Tib1 data. The detailed statistics of OLR data are listed in Table 3. We train the transformer model for this multilingual task. The modeling training settings are the same as the monolingual model.

Experimental Results: Table 4 shows that mixing the subset of AP-17 with NICT-Tib1 will bring slight improvement for every language except for the Uyghur language. This result empirically shows the usefulness of our dataset for multilingual ASR. More complex noisy conditionals and larger datasets will be investigated to verify this result.

4. CONCLUSION AND FUTURE WORK

This paper introduces the NICT-Tib1 (phase1) database, a new open-sourced database for the Lhasa dialect Tibetan language. We build ASR systems with this dataset. We further update the benchmark system with the multilingual setting. We will update the benchmark system with the state-of-the-art model in the near future. We believe our work will promote low-resourced language ASR research.

5. ACKNOWLEDGEMENTS

The work is partially supported by NICT international funding, and NICT tenure-track startup funding, Japan. We thank Siqing Qin, Lixin Pan, Prof. Longbiao wang, Prof. Jianwu Dang of Tianjin University, Dr. Xinhui Hu of RoyalFlush Inc., and Prof. Hao Huang of Xinjiang University for kindly supporting this work.

6. REFERENCES

[1] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1988.

[2] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[3] H. Wang, K. Khyuru, J. Li, G. Li, J. Dang, and L. Huang, "Investigation on acoustic modeling with different phoneme set for continuous Lhasa Tibetan recognition based on DNN method," in *Proc. APSIPA ASC*, 2016.

[4] J. Li, H. Wang, L. Wang, J. Dang, K. Khyuru, and G. Lobsang, "Exploring tonal information for lhasa dialect acoustic modeling," in *Proc. ISCSLP*, 2016.

[5] A. Graves and N. Jaitly, "Towards End-to-End speech recognition with recurrent neural networks," in *Proc. ICML*, 2014.

[6] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-End speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE-ASRU*, 2015, pp. 167–174.

[7] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, 2015.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE-ICASSP*, 2016.

[9] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected*

Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *arXiv preprint arxiv:1706.03762*, 2017.

[11] Lixin Pan, Sheng Li, Longbiao Wang, and Jianwu Dang, "Effective training end-to-end asr systems for low-resource lhasa dialect of tibetan language," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1152–1156.

[12] Siqing Qin, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan, "Improving low-resource tibetan end-to-end asr by multilingual and multilevel unit modeling," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, pp. 1–10, 2022.

[13] Siqing Qin, Longbiao Wang, Sheng Li, Yuqin Lin, and Jianwu Dang, "Finer-grained modeling units-based meta-learning for low-resource tibetan speech recognition," in *Proc. INTERSPEECH*, 2022.

[14] D. Wang, S. Ye, X. Hu, and S. Li, "The royalflush-nict system description for ap21-olr challenge (silk-road team, full tasks)," in *OLR2021 (oriental language recognition challenge)*, 2021.

[15] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, "An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model," in *Proc. INTERSPEECH*, 2021.

[16] Sheng Li, Xugang Lu, Chenchen Ding, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai, "Investigating radical-based end-to-end speech recognition systems for chinese dialects and japanese," in *INTERSPEECH*, 2019.

[17] Sheng Li, Chenchen Ding, Xugang Lu, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai, "End-to-end articulatory attribute modeling for low-resource multilingual speech recognition," in *Proc. INTERSPEECH*, 2019.

[18] Kak Soky, Sheng Li, Tatsuya Kawahara, and Sopheap Seng, "Multi-lingual transformer training for khmer automatic speech recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1893–1896.

[19] Yuqin Lin, Longbiao Wang, Jianwu Dang, Sheng Li, and Chenchen Ding, "End-to-end articulatory modeling for dysarthric articulatory attribute detection," in *Proc. IEEE-ICASSP*, 2020, pp. 7349–7353.

[20] *Tibeto-Chinese Lhasa Vernacular Dictionary (Tibetan)*, The Ethnic Publishing House, 1983.

[21] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE-ICASSP*, 2018.

[22] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," in *arXiv preprint arxiv:1806.05059*, 2018.

[23] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *arXiv preprint arxiv:1805.06239*, 2018.

[24] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *Proc. INTERSPEECH*, 2018.

[25] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. INTERSPEECH*, 2018.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018.

[27] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara, "Acoustic-to-word attention-based model complemented with character-level ctc-based model," in *Proc. IEEE-ICASSP*, 2018.

[28] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-Attention based End-to-End speech recognition with a deep CNN Encoder and RNN-LM," in *Proc. INTERSPEECH*, 2017.

[29] A. Baevski and et al., "wav2vec2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[30] A. Nautsch and et al., "The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding," 2019.