A study on speech recognition of Tibetan Amdo based on whisper

Like Ma[a], Guanyu Li*, Runyu Zhe[b]

Key Laboratory of Linguistic and Cultural Computing Ministry of Education（Northwest Minzu University), Lanzhou China

[a]y232430389@stu.xbmu.edu.cn, [*]xxlgy@xbmu.edu.cn,
[b]1036671790@qq.com

Abstract. In languages like Amdo Tibetan, which have a small speaker population and pose challenges in data collection, achieving high accuracy in speech recognition remains a considerable challenge. Whisper, a general-purpose speech recognition model developed by OpenAI, achieves near-human levels of accuracy and robustness by utilizing vast datasets for training. When the available Amdo corpus was utilized in this study, it was observed that after a brief period of fine-tuning, the Whisper model's recognition capabilities improved markedly. Initially unable to recognize Tibetan, the character error rate (CER) was reduced to 23.84% in the Whisper-base version post fine-tuning. Further improvements were noted in the Whisper-medium version, where the CER dropped to 9.31%. These findings highlight the Whisper model's substantial potential for recognizing low-resource languages and demonstrate the model's adaptability through fine-tuning for specific tasks. The study confirms that, despite limited data resources, targeted fine-tuning enables the Whisper model to achieve impressive recognition results in languages such as Amdo Tibetan.

Keywords: speech recognition; whisper; fine-tuning; Amdo Tibetan;

1 INTRODUCTION

The Whisper model emphasizes the impact of data volume on performance, choosing to use the Transformer[1] in an end-to-end architecture.

In automatic speech recognition (ASR) modeling, the efficacy of a model is contingent on both the quality and the quantity of the training data. In the case of specific domains or minority languages, the scarcity of datasets means that models are often unable to grasp the full complexity and diversity of the language. Conversely, training models with extensive datasets significantly enhances both performance and generalization capabilities, enabling the models to discern the fundamental rules of speech and language[2]. This study investigates how generalized speech recognition models like Whisper can be optimized for domain-specific performance through fine-tuning. Experimental results indicate that the model, when fine-tuned, adapts well to particular scenarios or requirements.

A. Haldorai et al. (eds.), *Proceedings of the 2024 3rd International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2024)*, Atlantis Highlights in Intelligent Systems 11,
https://doi.org/10.2991/978-94-6463-490-7_40

2 BACKGROUND

2.1 Whisper

The Whisper model developed by OpenAI demonstrated excellent generalization capabilities comparable to previous fully supervised learning results by training on 680,000 hours of multilingual and multitasking supervised datasets[3].To explore scaling properties, Whisper provides five different parameterized versions of the model to accommodate various application scenarios and resource constraints[4]. These variations allow users to select the most appropriate model based on their specific needs for accuracy, processing speed, and available resources.

2.2 LoRA

Low-Rank Adaptation (LoRA) is a technique that enhances resource efficiency by freezing the pre-trained model weights and incorporating trainable rank decomposition matrices into each Transformer layer[5]. This approach significantly reduces the number of trainable parameters. The core concept is depicted in Fig. 1: the original parameters W remain frozen, and only the matrices A and B are trained.

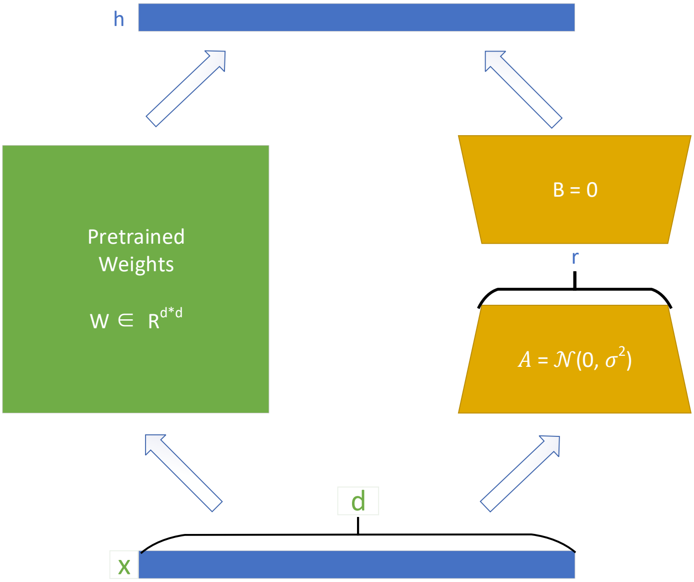

Fig. 1. LoRA fine-tuning

Matrix A reduces dimensionality, while matrix B increases it[6]. Initially, the weights of matrix A are set using a Gaussian function and the weights of matrix B are initialized to zero, ensuring that the pathway BA=0 at the start of training, thereby not affecting the initial model outcomes[7]. The outputs from the left and right pathways are combined to form new weight parameters that replace the original weights during inference. The details are shown in equation:

$$h = W_0X + \Delta W_x = W_0 \tag{1}$$

3 TRAINING

3.1 Amdo dialect dataset

The Amdo Tibetan dataset comprises a total of 31 hours of recordings made by 66 speakers, including 32 males and 34 females. The distribution of the data across these sets is detailed in Table 1 [8].

Table 1. Data Composition of Amdo Dialect

subset	Audio/h	male	female
train	25.41	27	27
dev	2.81	2	4
test	2.85	3	3

Far less time was spent using self-built datasets than common publicly available datasets, and while fine-tuning can significantly improve the model's recognition, sufficient data is essential to learn deeper information about a specific domain. In the future, we plan to expand the dataset by entering more compliant data on the one hand, and using language shifting and data augmentation to get more differentiated data on the other.

3.2 The token used in training

The Amdo Tibetan language comprises 35 consonants and 8 vowels. It uses a phonetic script where characters, known as syllables, are constructed from the Tibetan alphabet. The fundamental building blocks for these syllables are Tibetan letters, which combine according to a set of strict rules. As illustrated in Figure 2.

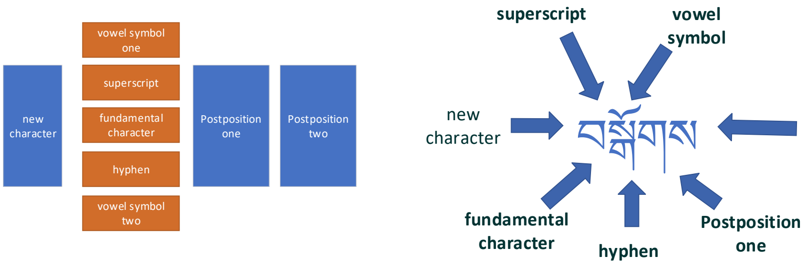

Fig. 2. Tibetan syllable structure and examples

we utilize bytecode as per the processing rules of the Whisper model. These rules involve converting the Tibetan alphabet into its corresponding bytecode. This bytecode is then transformed into characters based on the ASCII code. After conversion, this string is fed into the Whisper model's trained tokenizer, which further converts it into a sequence of numbers for training purposes[9,10]. For example, the Tibetan characters 'མ' are converted into bytecode represented as '\xE0\xBD\x98'. This bytecode is then transformed into characters based on the ASCII code table, resulting in the string 'à½Í,à½²'.After conversion, this string is fed into the Whisper model's trained tokenizer, which further converts it into a sequence of numbers for training purposes. The tokenizer used in this model contains 51,864 unique tokens, enabling the effective encoding of Tibetan script into a format suitable for machine learning applications.

Based on the above characteristics of Tibetan language, the study uses bytecode level token and chooses CER as a measure of model performance.CER is defined as the ratio of the minimum number of character operations required to convert the result of speech recognition to correct text to the total number of characters in the reference text. It is described below:

$$\mathrm{CER} = \frac{S_C+D_C+I_C}{N_C} \tag{2}$$

Where Sc is the number of character replacement errors, Dc is the number of character deletion errors, Ic is the number of character insertion errors, and Nc is the total number of characters in the reference text. The lower the value of CER, the higher the recognition accuracy of the model.

3.3 Fine-tuning the Whisper model

In Whisper, the target layer where LoRA needs to be injected is selected, which is usually the self-attention and feed-forward network (FFN) part of the Transformer layer. In the selected target layer, for the weight matrix W, introduce the low-rank matrices A and B. The product of these two matrices, BA, will simulate a small-scale update of the original weight matrix W. In this way, the weight matrix W will be updated to the original weight matrix W. In this way, the weight matrix W will be updated to the original weight matrix. While keeping the original weight matrix W unchanged, only A and B are updated during the training process.

For the input, the raw audio was segmented into 30-second segments and an 80-channel pairwise order-of-magnitude Mel spectrogram representation was computed as a feature representation using a 25-millisecond window on a 10-millisecond step. For labels, the addition of separators was used to form a complete sentence, which was encoded at the byte level. Finally loss calculation is performed as shown below:

$$\mathrm{L} = -\frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{C} y_{t,i}\log(\widehat{y_{t,i}}) \tag{3}$$

4 RESULTS

To investigate the effects of different parameter scales and separators on the Whisper model's performance in recognizing Amdo Tibetan speech, the study focused on fine-tuning the model across various configurations.

Table 2. Fine-tuning experiments at different scales and separators

Serial No	delimiter	Whisper model	CER
1	No delimiter	base	0.15649
2	<Space>	tiny	0.20422
3	“|”	tiny	3.96944
4	“་”	base	0.2402
6	“་”	small	0.11495
7	“་”	medium	0.09307
8	“་”	large-v2	0.10128

The results of the study are shown in Table 2, 1, 2, 3, 4 comparison results show that the selection of separator has a great impact on the model performance, 1 is better than the other experiments, but its prediction of something just Tibetan syllables without separator, if you use this method of identification in the future output process should be added to the language model to ensure that it can be correctly inserted into the separator; the selection of the space in the actual reasoning process, the tendency of the in the actual reasoning process, it tends to reason that meaningless garbled codes cannot be applied; selecting "|" in the prediction process, the Tibetan letters will swallow "|" making it impossible to be displayed, presumably due to the different encoding methods of Tibetan and "|". It is assumed that the reason is due to the different encoding methods of Tibetan and "|". For 5, 6 and 7, the model size does have an effect on the recognition rate, for 6 the effect is better than 7 because the Amdo Tibetan dataset is only about 31h, the reason for launching is that the data size is still too small to allow a larger model to learn the complete features, so it is necessary to select the appropriate size of the model for the dataset.

5 CONCLUSION

This paper focuses on the problem of poor speech recognition in low-resource languages, such as Amdo Tibetan. By implementing targeted fine-tuning on OpenAI's Whisper model, we investigate the use of a generalized speech recognition model to get satisfactory results for a particular problem under resource-constrained conditions.

Future experiments aim to enhance the Whisper model's performance for recognizing Amdo Tibetan by exploring various fine-tuning strategies. Key initiatives include integrating a language model to correct the recognized text, utilizing a larger-scale dataset for more robust training, selecting separators that are more effectively recognized by the model.

ACKNOWLEDGMENT

This research was financially supported by a Basic scientific research business project of central universities (31920220010).

REFERENCE

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://doi.org/10.5040/9781350101272.00000005 .
2. Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., & Norouzi, M. (2021). Speechstew: Simply mix all available speech recognition data to train one large neural network. arXiv preprint arXiv:2104.02133. https://arxiv.org/abs/2104.02133
3. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR. https://proceedings.mlr.press/v202/radford23a.html
4. Yang, H., Zhang, M., Tao, S., Ma, M., & Qin, Y. (2023, February). Chinese asr and ner improvement based on whisper fine-tuning. In 2023 25th International Conference on Advanced Communication Technology (ICACT) (pp. 213-217). IEEE. https://www.isca-archive.org/interspeech_2023
5. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685
6. Wang, X., Aitchison, L., & Rudolph, M. (2023). Lora ensembles for large language model fine-tuning. arXiv preprint arXiv:2310.00035. https://arxiv.org/abs/2310.00035
7. Liu, W., Qin, Y., Peng, Z., & Lee, T. (2024, April). Sparsely shared lora on whisper for child speech recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 11751-11755). IEEE. https://ieeexplore.ieee.org/abstract/document/10447004
8. Li, S. Y. (2023). Research on low-resource speech recognition based on transfer learning and language model fusion (Master's thesis, Northwest Minzu University). https://link.cnki.net/doi/10.27408/d.cnki.gxmzc.2023.000156.
9. Song, Z. H. (2023). Research on Tibetan Lhasa speech synthesis technology based on a completely end-to-end method (Master's thesis, Northwest Minzu University). https://link.cnki.net/doi/10.27408/d.cnki.gxmzc.2023.000290
10. Zheng, X., Zhang, C., & Woodland, P. C. (2021, December). Adapting GPT, GPT-2 and BERT language models for speech recognition. In 2021 IEEE Automatic speech recognition and understanding workshop (ASRU) (pp. 162-168). IEEE. https://ieeexplore.ieee.org/abstract/document/9688232