ManWav 1.0: The First Manchu ASR Model as the Milestone to Future Low-Resource ASR

Anonymous ACL submission

Abstract

This study addresses the widening gap in Automatic Speech Recognition (ASR) research between high resource and low resource languages, with a particular focus on Manchu, a severely underrepresented language. Manchu exemplifies the challenges faced by marginalized linguistic communities in accessing state-of- the-art technologies. In a pioneering effort, we introduce the first-ever Manchu ASR model ManWav 1.0, leveraging Wav2Vec 2.0 - XLSR. The results of the first Manchu ASR is promising, especially when our data augmentation method is employed. Wav2Vec 2.0 - XLSR fine-tuned with augmented data demonstrates a 2%p drop in CER and 13%p drop in WER compared to the same model fine-tuned with original data. This advancement not only marks a significant step in low resource ASR research but also incorporates linguistic diversity into technological innovation.

1 Introduction

The landscape of Automatic Speech Recognition (ASR) research has centered around high resource languages, prominently exemplified by the extensive focus on languages like English. This concentrated attention on high resource languages has inadvertently deepened the divide between research advancements. While research on English ASR encompasses diverse linguistic variations, including accented and noised speech, the same cannot be said for many low resource languages, though a few basic research such as Safonova et al. (2022) and Zhou et al. (2022) exist. Astonishingly, not a single basic ASR model has been developed for Manchu to date, highlighting a critical void in linguistic inclusivity within the realm of ASR technology.

This paper sets out to address this significant gap by undertaking the ambitious task of developing the inaugural Manchu ASR model. The endeavor is underscored by the scarcity of linguistic resources, prompting us to collect all existing Manchu audio data from Kim et al. (2008) in one channel. We try to maximize the cross-lingual capabilities of Wav2Vec 2.0 - XLSR (Conneau et al., 2020) by fine-tuning the model with Manchu audio data. The performance of Manchu ASR model is further enhanced through data augmentation. This approach reflects a strategic adaptation to the challenges posed by the limited availability of annotated data for low resource languages, as we strive to contribute not only to the nascent field of Manchu ASR research but also to the broader discourse on linguistic inclusivity in cutting-edge technology.

The contributions of this study are as follows:

- Collecting Manchu audio data and correcting corresponding transcriptions
- Augmenting the limited amount of Manchu data up to 4 times
- Developing the very first Manchu ASR model

2 Manchu Language

The Manchu language, a member of the Tungusic linguistic family, has its roots among the Manchu people of Northeast China and boasts a significant historical role as the official language of the Qing dynasty (1644-1912). Presently, the language confronts a dire state of endangerment, officially denoted a dead language.

There have been some efforts to employ technological solutions in the preservation and revitalization of Manchu. These endeavors include the Manchu spell checker(You, 2014), *Mergen*: Manchu-Korean machine translation(Seo et al., 2023), and Manchu NER/POS tagging models[1]. However, due to the paucity of data, studies above face challenges and no ASR model has been yet developed.

[1]https://github.com/sanajlee/Manchu-NLP

3 Data

3.1 Materials

This study leverages Colloquial Manchu data provided by Kim et al. (2008), in which Colloquial Manchu data is gathered as part of ASK REAL project (Altaic Society of Korea, Researches on Endangered Altaic Languagess(Choi et al., 2012)). This audio data represents the dialect of Shanjiazi village, Located in the Youyi Dowoerzu Manzu Ke'er-kezizu township, Fuyu county, Heilongjiang Province.

The recording took place from February 7th to 14th, 2006 in Qiqihar, Heilongjiang Province, with Mr. Meng Xianxiao (73 years old at that moment). Though Chinese being his first language, Mr. Meng Xianxiao sufficiently served as the consultant as he acquired a comprehensive understanding of the Manchu language by the age of 12.

The data we use in the study is the recordings of the basic conversational expressions and the sentences for grammatical analysis. The length of each recordings is 1,944.44 seconds and 3,513.90 seconds, in total of 5,458.34 seconds. Corresponding transcriptions are basically provided by Kim et al. (2008), but slight adjustments are added in order for better alignment with the consultant's authentic language habits. These habits include stuttering and interjections.

3.2 Transcription

The phoneme transcription system in this study is based on Kim et al. (2008). While it shares similarities with the International Phonetic Alphabet (IPA), our system incorporates some distinctions. Specifically, /b, d, g/ represent voiceless unaspirated stops, and /p, t, k/ denote voiceless aspirated stops. Notably, Colloquial Manchu lacks voiced stops, making this transcription system more practical than using diacritic /ʰ/ to indicate aspiration. Next, /ǰ, č, š/ denote voiceless palatal sounds. In IPA system, corresponding sound symbols are [ɟ, ç, ɕ]. But /ǰ/ is not voiced unlike [ɟ], and /č/ is the aspirated sound, [čʰ]. Some examples can be found in Table 1.

Transcription	Translation
miŋ ənjə bitk səwə.	My mother is a teacher.
došən ǰo.	Come on in.

Table 1: Examples of our transcription and corresponding translation.

3.3 Data Augmentation

The scarcity of speech datasets from native Manchu speakers presents a significant challenge, necessitating the adoption of various data augmentation methods. Audio data augmentation methods used to simulate different acoustic environments include:

- **Additive noise**: Adding background noise to the audio samples.
- **Clipping**: Involves cutting short the audio signals.
- **Reverberation**: Applying reverberation effects.
- **Time dropout**: Randomly removing segments of the audio.

By implementing the above techniques through WavAugment[2] provided by Kharitonov et al. (2020), we successfully expand the dataset by 100% respectively, to a total of 400%, significantly enriching the available data ASR model training. Notable is the fact that data augmentation is implemented after the separation of train and test data, in order for more reliable test results. The size of data before and after augmentation is described in Table 2.

Before Augmentation	Duration
train	4,898.71s
test	559.63s
After Augmentation	**Duration**
train	19,594.85s
test	559.63s

Table 2: The duration of audio files(.wav) in seconds before and after augmentation.

4 Experiment

4.1 Models

Wav2vec2-large-xlsr-53(Conneau et al., 2020) is utilized as the base model to evaluate the influence of data augmentation. Wav2vec2-large-xlsr-53 is a multilingual ASR model from Meta AI pre-trained with 53 languages. A Wav2vec2-large-xlsr-53 model is fine-tuned in two different types of data, leading to two separate fine-tuned models: one with original Manchu data, and the other

[2]https://github.com/facebookresearch/WavAugment

with augmented Manchu data. We name the model trained with augmented data ManWav 1.0. The fine-tuning process is conducted through HuggingSound(Grosman, 2022).

4.2 Experimental Setup

Our experiments are conducted using an NVIDIA A100 GPU. We fine-tune our models with learning rates ranging from 1e-5 to 5e-5, batch sizes of 4, 8, 16, 32, and dropout rate of 0.1. Each model is trained for up to 5 epochs, providing a balanced approach to assess performance and efficiency under various configurations.

5 Result and Discussion

5.1 Result

We use Character Error Rate (CER) and Word Error Rate (WER) as evaluation metrics. CER assesses the accuracy of character transcription, while WER measures the correctness of word recognition. Scores closer to 0 represent better performances in both metrics. WER and CER are the most common and essential metrics in gauging the overall performance of ASR systems.

The experimental results prove the significance of data augmentation in fine-tuning the pre-trained ASR model. As depicted in Table 3, using augmented data at the training stage clearly improves the performance of Wav2vec2-large-xlsr-53, specifically dropping CER by 0.02 and WER by 0.13, indicating the effectiveness of our data augmentation methods described in Section 3.3. Further, inference results of wav2vec-based models trained on other languages can be found in Table 4.

Moreover, Table 5 shows the promising capabilities of our model in Manchu speech recognition tasks. The achieved accuracy is particularly noteworthy given the limited availability of Manchu speech data and considering that Wav2vec2-large-xlsr-53 is not initially pre-trained on Manchu. These outcomes highlight the adaptability and potential of our approach in processing languages with scarce resources.

Data Augmentation	CER	WER
before	0.13	0.44
after	**0.11**	**0.31**

Table 3: The performance of wav2vec2-large-xlsr-53 each trained with data before and after augmentation.

5.2 Linguistic Analysis

The discrepancies between the inference results and the original labels in the test data can be broadly classified into three scenarios: (1) confusion involving /ə/, (2) confusion related to nasal sounds in word-final positions, and (3) confusion between /w/ and /x/.

First, there are some uncaptured or mismatched /ə/ sounds in the inference results, particularly in word-final or between sonorants (e.g., /l/) and stops. This occurs because /ə/ can be neutralized with other vowels or even deleted, posing challenges in accurate transcription.

Secondly, nasal sounds /n/ and /m/ in word-final positions are frequently overlooked in the inference results. This could be attributed to the nature of nasal sounds, as they tend to be fused with subsequent vowels, resulting in nasalized vowels, or they may be omitted altogether.

Lastly, there is a frequent confusion between intervocalic /w/ and /x/ in the inference results. Given that /w/ is the labial approximant and /x/ is the palatal approximant, it can be noted that these two sounds occupy distinct articulatory positions. However, there is no equivalent unvoiced sound for /w/, and discerning the voicing of approximants becomes challenging when they are in intervocalic positions.

The above three types of mismatch and corresponding examples are elaborated in Table 6.

6 Related Work

6.1 ASR research in low-resource languages

There exist some endeavors to apply ASR to low-resource languages. For example, Safonova et al. (2022) collect a speech dataset in the Chukchi language and train an XLSR model. Similarly, Qin et al. (2022) improve low-resource Tibetan ASR while Jimerson and Prud'hommeaux (2018) introduce a fully functional ASR system tailored for Seneca, an endangered indigenous language of

Model	CER	WER
ManWav 1.0	**0.11**	**0.31**
wav2vec2-base-960h	0.68	1.12
wav2vec2-base-100h	0.67	1.15
wav2vec2-large-960h-lv60-self	0.59	1.01
wav2vec2-large-lv60	1.66	1.0

Table 4: Inference results on Manchu test data of other wav2vec-based models trained on English

Model Prediction	Actual Transcription
mim bo də ilan bo d bi mim bo ilan bo d bi əltə tələm tələčəl dulilə ajlə lod man njam bi jə tələ amba njam wakə tələ amba njam wakə	mim bo də ilan bo bim mim bo ilan bo bim əltə tələm tələčəl dulilə ailə lod man njam bi jə tələ amba njam wakə tələ amba njam wakə

Table 5: Examples of inference results from fine-tuned wav2vec2-large-xlsr-53.

Mismatch Types	Examples
(1) ə / __# or R__C	amə : am, duləkə : dulkə
(2) n, m / __#	ilan : ila, jom : jo
(3) w : x / V__V	šaxulo : šawulo, indaxo : indawə

Table 6: Observed mismatch examples from the inference results written in phonological notations. R refers to sonorants, C consonants, and V vowels.

North America. Simultaneously, Singh et al. (2023) propose an effective self-training approach capable of generating accurate pseudo-labels for unlabeled low-resource speech, particularly for the Punjabi language. In addition, researchers have delved into exploring training strategies that lead to more efficient data utilization for low-resource speech recognition, as highlighted by Zhou et al. (2022). Additionally, there has been research dedicated to investigating data augmentation methods to enhance ASR systems designed for low-resource languages (Bartelds et al., 2023).

6.2 Wav2Vec 2.0

Wav2Vec 2.0 (Baevski et al., 2020) is a state-of-the-art speech recognition model developed by Facebook AI[3]. The core innovation of Wav2Vec 2.0 lies in its ability to effectively capture the contextual information in speech through its Transformer-based architecture. Wav2Vec 2.0 leverages self-supervised training, allowing the training of an ASR model with a minimal amount of labeled data, provided there is an ample supply of unlabeled data. The performance of a pretrained Wav2Vec 2.0 model using untranscribed data demonstrates notable accuracy when fine-tuned on a relatively modest quantity of transcribed data. Noteworthy is Wav2Vec 2.0's effectiveness not only in capturing diverse dialects but also in accommodating various languages. XLSR models (Conneau et al., 2020) are multilingual Wav2Vec 2.0 checkpoints particularly pre-trained on 53 languages and fine-tuned for Connectionist Temporal Classification(CTC) speech recognition. CTC is a technique used in encoder-only transformer models such as Wav2Vec 2.0(Baevski et al., 2020), HuBERT(Hsu et al., 2021) and M-CTC-T(Lugosch et al., 2022).

7 Conclusion

ManWav 1.0 is a significant contribution to the field of ASR by focusing on Manchu, an extremely low resource language often overlooked in linguistic technology. The development and implementation of the first-ever Manchu ASR model, utilizing Wav2Vec 2.0 - XLSR, marks a groundbreaking step in addressing the disparities between high resource and low resource languages in ASR research. This achievement not only highlights the need to support underrepresented linguistic communities but also showcases the potential of cutting-edge ASR technology to make a substantial impact in bridging these gaps. The experiments of this research show the effectiveness of several data augmentation methods in low resource scenarios, namely additive noise, clipping, reverberation, and time dropout.

8 Future Work

In further studies, we will focus on enhancing the inference quality with the help of a language model. The addition of a decoder to an ASR model is known to boost the inference performance(Karita et al., 2019; Zeyer et al., 2019). A decoder model such as GPT model is expected to be well trained with Manchu data. This belief stems from the existence of a BERT-base model[4] successfully pretrained leveraging Manchu text data from Seo et al. (2023).

This future research will not only improve the Manchu ASR model but also contribute to a deeper understanding and preservation of such underrepresented languages. Furthermore, the successful development of a Manchu-specific decoder could serve as a model for similar advancements in other low-resource languages, thereby fostering a more inclusive and diverse representation in the field of speech recognition technology.

[3] https://ai.meta.com/

[4] https://github.com/seemdog/manchuBERT

Limitations

The primary constraint of this research lies in the scarcity of Manchu audio data. As the only audio data used in this research consists of only Colloquial Manchu, utilizing ManWav1.0 in other domains would not show optimized performances, given that ASR models are usually heavily domain-dependent.

Because Manchu is a dead language, without native Manchu speakers, generation of new Manchu audio data is an insurmountable challenge. Further, only a small proportion of Manchu audio data is transcribed manually, compounding more difficulty in training Manchu ASR models.

Ethics Statement

The project paves the way for further innovations in the field and emphasizes the importance of inclusivity in technological advancements, ensuring that the benefits of state-of-the-art technologies are accessible to all linguistic groups, regardless of their resource status. To support the further ASR studies on endangered languages, we plan to release our Manchu ASR model in public shortly.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation.

Wonho Choi, Hyunjo You, and Juwon Kim. 2012. The documentation of endangered altaic languages and the creation of a digital archive to safeguard linguistic diversity. *International Journal of Intangible Heritage*, 0(7):103–111.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

Jonatas Grosman. 2022. HuggingSound: A toolkit for speech-related tasks based on Hugging Face's tools. https://github.com/jonatasgrosman/huggingsound.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2020. Data augmenting contrastive learning of speech representations in the time domain.

Juwon Kim, Dongho Ko, Chaoke D. O., and Boldyrev B. V. Han Youfeng, Piao Lianyu. 2008. *Materials of Spoken Manchu*. Seoul National University Press.

Loren Lugosch, Tatiana Likhomanenko, Gabriel Synnaeve, and Ronan Collobert. 2022. Pseudo-labeling for massively multilingual speech recognition.

S. Qin, L. Wang, S. Li, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan. 2022. Improving low-resource tibetan end-to-end asr by multilingual and multilevel unit modeling. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022.

Anastasia Safonova, Tatiana Yudina, Emil Nadimanov, and Cydnie Davenport. 2022. Automatic speech recognition of low-resource languages based on chukchi.

Jean Seo, Sungjoo Byun, Minha Kang, and Sangah Lee. 2023. Mergen: The first manchu-korean machine translation model trained on augmented data.

Satwinder Singh, Feng Hou, and Ruili Wang. 2023. A novel self-training approach for low-resource speech recognition.

Hyun-Jo You. 2014. A manchu speller: With a practical introduction to the natural language processing of minority languages. *Altai Hakpo*, 24:39–67.

Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schluter, and Hermann Ney. 2019. A comparison of transformer and lstm encoder decoder models for asr. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15.

Zhikai Zhou, Wei Wang, Wangyou Zhang, and Yanmin Qian. 2022. Exploring effective data utilization for low-resource speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8192–8196.