Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070302
ISSN 2371-8838 Vol. 7 Num. 3

Tibetan Speech Recognition Based on Wav2vec Feature

Zixi Yan, Guanyu Li*, Senyan Li

Northwest Minzu University, Lanzhou, Gansu, 730000, China
**Corresponding author*

Keywords: Tibetan, wav2vec2, speech recognition, kaldi, low resource

Abstract: Speech recognition tasks for small languages such as Tibetan language have been unable to achieve the same results as those for large languages. In this paper, the wav2vec2 model is introduced into the traditional model to extract features and improve the effect of Tibetan speech recognition. In this paper, the kaldi tool was used to construct a speech recognition system for Tibetan language, and the wav2vec2 model was used as the feature extractor to replace the traditional mfcc features. The improvement effect of the front-end model and traditional speech recognition model on speech recognition of minority languages such as Tibetan wa comparatively anayzed, and the effectiveness of wav2vec2 model in non-pre-trained languages was verified. Finally, the recognition efficiency of the proposed model on per and wer was increased by 2.92% and 5.24% respectively as compared with the baseline system.

1. Introduction

With the emergence of various model structures in the study of large languages such as English and Chinese in the field of artificial intelligence, speech recognition research related to Tibetan as a low-resource language has been developing. In 2007, Liu Jingping et al. [1] proposed isolated word recognition based on dynamic time warping [2]. With the development of Hidden Markov Model (HMM) [3], hidden Markov model is also introduced in Tibetan speech recognition research. At the beginning, it is isolated word recognition based on hidden Markov model, for example, Pei Chunbao [4] realized Tibetan 1-9 digits recognition based on HMM. Later, from 2012 to 2015, more researchers built Tibetan continuous speech recognition systems based on HMM, such as the current customized continuous speech recognition system built by Li Guanyu et al. [5] However, since these researchers mostly used their own Tibetan language data sets, the recognition rate between them could not be compared.

Self-supervised (SSL) speech representation based on pre-training has achieved great success in many speech tasks. The task of speech recognition in small languages such as Tibetan has been unable to achieve the same effect as that of large languages, which is largely due to the lack of training corpus. The pre-training model uses a large amount of corpus for self-supervised pre-training and learns general knowledge of speech. Taking the pre-training model as the upstream model, the knowledge of the pre-training model can improve the training burden of the downstream model, which is very helpful for the speech recognition of small languages that lack corpus. In this paper, the pre-training model of wav2vec2 [6-8] was used to extract speech features and replace traditional features in the speech recognition system, so that the Tibetan language recognition system could make use of the

relevant knowledge learned in the pre-training model to improve the effect of the Tibetan language recognition system.

In this paper, the Tibetan speech recognition system was also built based on kaldi and taken as the baseline system. After that, wav2vec2 model was introduced as the feature extraction scheme to replace mfcc features, and the effect of pre-trained wav2vec2 model applied to Tibetan speech recognition was studied.

2. Method and Experiment

2.1 Wav2vec2 Model

Wav2ve2 model consists of a multi-layer convolutional feature encoder f: X→Z. It takes the raw audio X as input and outputs the underlying voice representing z1... , zT time step, and then inputs them into a transformer layer g: Z→C to build the representation c1... , cT and captures information from the entire sequence. The output of the feature encoder is discretized as qt, and a quantization module Z→Q is used to represent the target under self-supervision. In contrast to vq-wav2vec, the wav2vec2 model builds context representation on continuous speech representation, and captures end-to-end dependencies on the entire sequence of potential representations.

2.2 Tibetan Speech Recognition System Based on Kaldi

Kaldi is a mature speech recognition system framework, based on which we can easily build various speech recognition systems. In this paper, kaldi was used to build a Tibetan speech recognition system based on hmm. The structure of this model is similar to the continuous speech recognition system built by Li Guanyu [5], which requires training of acoustic model and ngram language model respectively, and finally decoding and recognition. The speech features used in the baseline model are mfcc features.

2.3 Feature Displacement

The baseline Tibetan speech recognition system uses 13-dimensional mfcc features, while the pre-trained wav2vec2 system extracts features of 1024 dimensions at each layer, which is not suitable for the traditional speech recognition system. Building traditional speech recognition system by directly using features with too high dimensions will result in an oversized model, and the convergence of an oversized model is not easily trained, which is a huge defect for speech recognition of small languages that lack corpus. Therefore, pca dimension reduction technology was used in this paper to reduce the dimension of the features extracted by wav2vec2 system, so as to adapt to the traditional speech recognition system. Different dimensionality reduction techniques such as pca have their own characteristics and focus on the retention of high-dimensional spatial information. Therefore, different dimensionality reduction techniques need to be tested separately to obtain the best dimensionality reduction effect.

Pca dimensionality reduction techniques require a number of samples to calculate the final dimensionality reduction projection matrix. In general, training set samples are used to calculate dimensionality reduction projection matrix in the training stage, and this projection matrix is adopted as the final projection matrix. In the test phase, the projection matrix calculated in the training phase is used to reduce the dimension of wav2vec2 features of the test data. The calculation of the projection matrix consumes a lot of hardware resources, so part of the training set samples are also used in the calculation of the projection matrix, based on which the dimensionality reduction of the remaining training set and test set data is calculated.

2.4 Data Set

The Tibetan language data set used in this paper was the Tibetan language corpus recorded by the laboratory.

2.5 Experiment

In this experiment, 6.69mb Tibetan text was used to train the ngram language model [9], and the training set data was used to train the acoustic model, and finally the test set was used to test the model effect. The acoustic model training was conducted under two situations. The tools provided by kaldi was used in baseline system to extract the mfcc features, while the wav2vec2 model was used in the contrast experiment to extract the features. By using the model XLSR-53 [10] published online for pre-training in multiple languages, wav2vec2 model can extract 24 layers of different features, and the effects in different tasks are also different. The 24 layers of features were used to replace the mfcc features and train the acoustic model. Finally, the speech recognition experiment was conducted, and the feature of the layer with the best effect among the 24 layers of features was selected as the experimental result, as shown in table 1.

3. Result and Analysis

Table 1: Speech recognition results

System	Per	Wer
Baseline	16.09	23.89
XLSR-53 + pca dimension reduction (15 layers)	13.17	18.65

As shown in Table 1, replacing mfcc features in hmm speech recognition system with features extracted from wav2vec2 model can greatly improve the speech recognition of low resources. In this experiment, the speech recognition on per and wer were 2.92% and 5.24% higher than that of baseline system, respectively.

In this paper, Wav2vec2 model was used as the pre-training model. Tibetan corpus was not used in the pre-training process, but the model was applied to Tibetan speech recognition, still achieving great improvement effect. This shows that the wav2vec2 model does learn information during pre-training self-supervised learning. In the hierarchy, it should be above the acoustic information, but below the linguistic information. Firstly, the effect of this model feature is better than the mfcc feature, indicating that the feature should contain more information that are easier to be processed than the mfcc. However, without using Tibetan language for training, the features extracted by this model still improved the recognition effect of Tibetan speech recognition system. Therefore, the model should not contain too much information related to the language itself, because different languages have conflicting parts in the language structure. This should be due to the fact that XLSR-53, the open wav2vec2 model, uses the corpus of more than 53 different languages when training.

4. Conclusion

When the wav2vec2 model with multi-language corpus for pre-training is used as the front-end system to replace the mfcc in traditional speech recognition for feature extraction, great improvement can be achieved even though the wav2vec2 model does not use the corpus of the target language in the pre-training stage. This proves that wav2vec2 model is an excellent model to improve the speech recognition effect of low-resource languages.

Acknowledgments

This research was financially supported by "the Fundamental Research Funds for the Central Universities" (31920220010).

References

[1] Liu Jingping and Dexi Jiacuo, Design of ando-tibetan consonant recognition, Research on Ethnic Language Information Technology -- The 11th National Ethnic Language Information Symposium. 2007: p. 11-15.
[2] Müller and Meinard, Information retrieval for music and motion. 2007: Springer Berlin Heidelberg. 69-84.
[3] Rebello Sinda and Y.H. Y, An integrated approach for system functional reliability assessment using Dynamic Bayesian Network and Hidden Markov Model. Reliability Engineering & System Safety, 2018: p. 124-135.
[4] Pei Chunbao, Research on Tibetan Speech Recognition Technology based on Standard Lhasa language, 2009, Tibet University: Lhasa.
[5] Li Guanyu and Meng meng, Study on Acoustic Model for Continuous Large Vocabulary Speech Recognition of Tibetan Lhasa Dialect. Computer Engineering, 2012: p.189-191.
[6] Baevski A., et al., wav2vec 2.0 A Framework for Self-Supervised. 2020.
[7] Baevski A., S. Schneider and M. Auli, vq-wav2vec- Self-Supervised Learning of Discrete Speech Representations. 2020.
[8] Conneau A., et al., {Unsupervised Cross-lingual Representation Learning for Speech Recognition}. arXiv e-prints, 2020: p. arXiv:2006.13979.
[9] Yin Chen and Wu Min, A review of N-gram model. Computer Systems & Applications, 2018. 27(10): p.33-38.
[10] Xu Q., A. Baevski and M. Auli, Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. arXiv e-prints, 2021: p. arXiv:2109.11680.