The TNT Team System Descriptions for OpenASR21

Kai Tang†, Jinghao Yan, Jian Kang, Shen Huang*, Pengfei Hu[1]$
TEG AI/CSIG AI(*), ASR oteam
Tencent Inc, Beijing 100193, China
alanpfhu@tencent.com

Jing Zhao†, Haoyu Wang, Jinpeng Li, Shuzhou Chai, Guan-Bo Wang, Guoguo Chen, Wei-Qiang Zhang[1]$
Department of Electronic Engineering
Tsinghua University, Beijing 100084, China
wqzhang@tsinghua.edu.cn

Abstract—This paper presents our architecture and a series of experiments for OpenASR21. We describe the systems in the constrained condition, constrained-plus condition and unconstrained condition, and our post evaluation analyses. The systems in constrained condition are nearly the same as that in last year. For constrained-plus condition, pre-training models and system fusions are incorporated. We pre-train our conformer encoders using wav2vec 2.0 pre-training method, which are called Ch-w2v-conformer. For unconstrained condition our end to end ASR systems with conformer in optimized loss and long sequence encoders are adopted. We also adapt this challenging PSTN conditions using public availably data in shape of wideband dictated speech with similar accent, respectively. Finally, series of systems are submitted for this challenge. The WER of our submitted system for constrained condition is about 0.4, and for constrained-plus condition, the result is about 0.3~0.4, which are about 0.05 lower than that in constrained condition. For unconstrained condition, most of the languages could be below 0.4, in terms of WER. We do NOT manage to submit all the systems so left results are summarized in this report.

Keywords—automatic speech recognition, low resource languages, OpenASR21, speech pretraining

I. Introduction

Due to the lack of speech data, language script, lexicons, building an applicable ASR system for low resourced language is very challenging. The goal of the OpenASR21 Challenge is to assess the state of the art of ASR technologies under low-resource language constraints. It consists of performing ASR on audio datasets in up to fifteen different low resource languages, producing the recognized written text. For constrained condition, participants are only given 10 hours subset of labelled acoustic data but extra text data is unlimited. For constrained-plus condition, except the given 10 hours labelled acoustic data, pretrained models can be finetuned by the given 10 hours training data. For unconstrained condition, teams may use speech data outside of the 10-hour subset marked for the constrained condition for the language being processed, as well as additional publicly available speech and text training data from any languages. The evaluation dataset is provided a week before the system submission deadline.

We participate in 2 languages in Constrained condition, 3 languages in Constrained-plus condition and Unconstrained condition, i.e. Cantonese, Mongolian and Kazakh.

II. Constrained System

For our hybrid acoustic model, we propose the CNN-TDNN-F-A network as the essential part, which is trained with lattice-free maximum mutual information (LF-MMI) criterion [1]. The model introduces self-attention mechanism [2] to the combination of CNN and TDNN-F [3] in order to learn more positional information from the input.

Since the major challenge is low-resource condition, various kinds of data augmentation methods are combined to get additive improvement, such as speed perturbation [4], volume perturbation [4]. These are proved to be effective to ASR performance especially under low-resource condition.

Besides, systems' diversity is important for the final fusion. We have trained more than four systems of each language to make further use of the diversities of single system by system fusion. The systems in constrained condition are nearly the same as that in last year. The details are in [5].

The main workflow is illustrated in Fig. 1, which consists of pre-processing, data augmentation, training, decoding and system fusion roughly. Our systems' performance of the

$ The corresponding author is Wei-Qiang Zhang and Pengfei Hu

† equal contribution

constrained condition on the evaluation set is shown in Tab.1, which are released by NIST OpenASR scoring server.

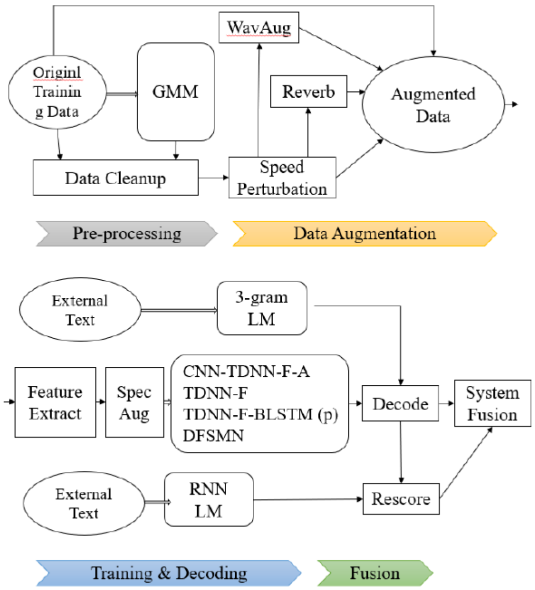

Figure 1: Workflow of the ASR systems: The whole system process can be roughly divided into data cleanup, pre-processing, data augmentation, training, decoding and system fusion.

TABLE I WER OF ASR SYSTEMS ON DEV AND EVAL SET (CONSTRAINED)

	WER on DEV	**WER on EVAL**
Cantonese	*0.412*	*0.402*
Mongolian	*0.461*	*0.378*

III. CONSTRAINED-PLUS SYSTEM

A. Unsupervised speech pretrain

We pre-train our conformer encoders [6] using wav2vec 2.0 pre-training method [7,8], which we called Ch-w2v-conformer. The original pre-training works take raw waveforms as input. Unlike these works, we use MFCC features as inputs. The Conformer encoder are split into a "feature encoder" and a "context network" naturally. The former one, i.e. “feature encoder”, consists of convolution subsampling block. The latter one, i.e. "context network" is made of a stack of conformer blocks. There are two 2D-convolution layers in a convolutional subsampling block. The time stride of convolutional subsampling blocks is 3, so the length of the input feature sequence becomes one sixth. The encoded features from the convolutional subsampling block are fed into the conformer context network to make context vectors. Meanwhile, the encoded features are passed through a quantization layer to produce target context vectors. Wav2vec 2.0 pre-training algorithm[7] optimizes the contrastive loss between the context vectors from the masked positions and the target context vectors. This procedure is presented in Fig.2 below.

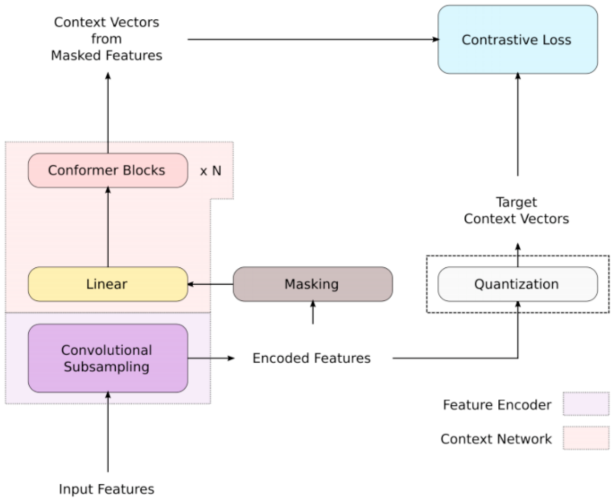

Figure 2: The Ch-w2v-conformer pre-training architecture

18 stacked conformer blocks are used in our Ch-w2v-conformer work. Each block contains four modules: a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module. To speed up training procedure, the original relative positional encoding is removed and the order of convolution module and multi-head self-attention module is swapped. Fig. 3 shows the architecture of convolutional feature encoder and conformer block.

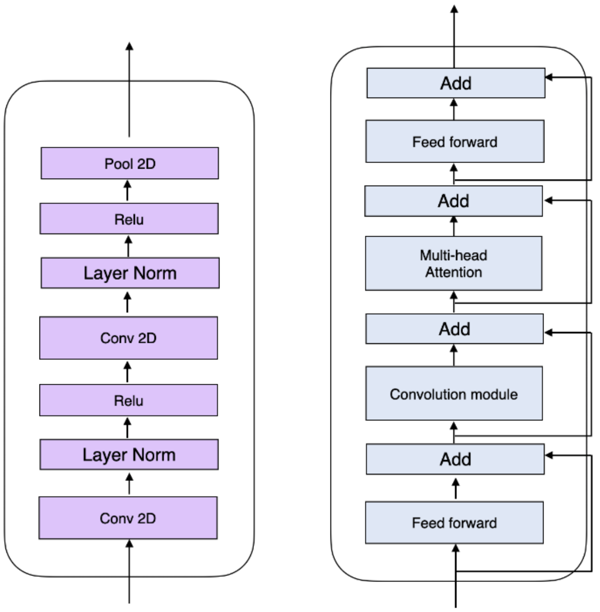

Figure3: The left part is convolutional feature encoder block. The right part is conformer block.

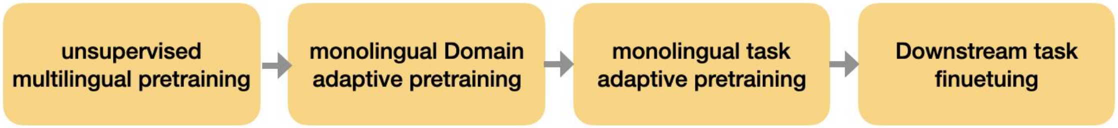
unsupervised multilingual pretraining
monolingual Domain adaptive pretraining
monolingual task adaptive pretraining
Downstream task finuetuing

Figure4: continue pre-training and fine-tuning process of pre-training model

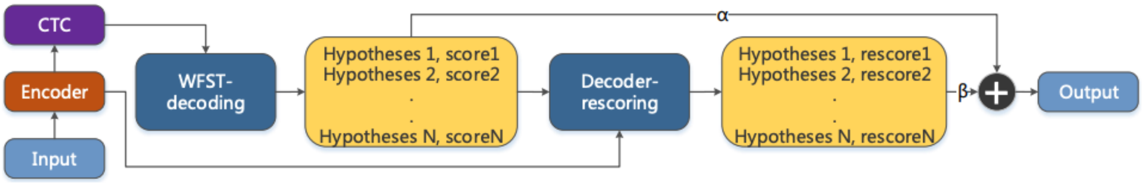

Figure5: the two-pass decoding framework[24]

The Ch-w2v-conformer model uses following datasets to pretrain: ISML datasets (6 languages, 80k hours): internal dataset contains Chinese, Cantonese, Tibetan, Inner Mongolian, Inner Kazakh and Uighur. Babel datasets (17 languages, 2k hours): This is a multilingual corpus of conversational telephone speech from the IARPA program. We adopt the same data setup as XLSR model [9] and pretrain on 17 languages, consisting of Assamese, Bengali, Cantonese, Cebuano, Georgian, Haitian, Kazakh, Kurmanji, Lao, Pashto, Swahili, Tagalog, Tamil, Tok pisin, Turkish, Vietnamese and Zulu.

Our pre-trained model is trained using data of 16kHz sampling rate. Unfortunately, most of the OPENASR DEV data is recorded in telephone channel with a sampling rate of 8kHz. In order to obtain better downstream task performance, we use the multiphase adaptive pretraining method similar to NLP tasks [10]. First, we continue pretraining Ch-w2v-conformer on a corpus of unlabeled domain data of the corresponding language. Second, adapting to the unlabeled data of specific task (task-adaptive pretraining) improves performance even after domain-adaptive pretraining. Fig. 4 shows the process of pretraining and fine-tuning our models. This model is open source in [11].

We also use Facebook's open source XLSR model [9] to finetune the competition task. The XLSR model uses 53 languages for multilingual pre-training consisting of MLS dataset(8 languages), Commonvoice(36 languages) and Babel(17 languages). It is discovered that the XLSR model and Ch-w2v-conformer model have very complementary effects at the system fusion phase.

B. *Downstream finetune*

After pretraining, we build end to end conformer based ASR system using the given 10 hours labelled acoustic data in constrained condition. We use hybrid CTC-Attention loss to finetune the Ch-w2v-conformer model [12]. Both CTC loss and attention loss are used to train end to end ASR system. This multitask training algorithm helps the model converge faster and achieves better performance.

C. *Language models and decoding*

During decoding stage, we use a two-pass decoding framework based on CTC-Attention architecture [13]. During CTC decoding at the first-pass stage, we train a 4-gram model using babel text and build a WFST framework for decoding. Then, the N-best hypotheses are rescored by the attention-decoder at the second pass stage. Fig.5 shows the proposed decoding framework.

D. *System fusion*

As we know, the results of systems with different data augmentations and network architectures have complementary effects. So we select some systems to fuse in order to achieve better performance. After recognition results of single system are obtained, ROVER is adopted. ROVER is a post-recognition process that models the output generated by multiple ASR systems as independent knowledge sources. ROVER proves to be effective in reducing word error rate [14]. Tab. 2 shows the performance of single systems and fused system. Tab. 3 shows the performance of the evaluation set in constrained-plus condition, which are released by NIST OpenASR scoring server.

TABLE 3 WER OF ASR SYSTEMS ON DEV AND EVAL SET (CONSTRAINED-PLUS)

	WER on DEV	*WER on Eval*
Cantonese	*0.326*	*0.337*
Mongolian	*0.341*	*0.378*
Kazakh	*0.345*	*0.428*

TABLE 2 WER OF SINGLE ASR SYSTEMS ON DEV SET (CONSTRAINED-PLUS)

System	*WER*		
	Cantonese	*Mongolian*	*Kazakh*
Constrained Hybrid system	*0.412*	*0.455*	*0.464*
Ch-W2v-conformer	*0.366*	*0.433*	*0.434*
XLSR	*0.373*	*0.439*	*0.421*
Rover *Fusion*	*0.347*	*0.410*	*0.387*

IV. UNCONSTRAINED SYSTEM

For unconstrained condition, due to the time limit, we only participate in Cantonese, Mongolian and Kazakh. The main fused system is roughly the same with constrained system, except that end to end (e2e) ASR training, hybrid bandwidth acoustic model, language optimization and hybrid-e2e fusions are explored additionally.

A. End to end system

The end to end ASR in our system is based on recent conformer [6] structure. As transformer, the conformer model includes two parts: encoder and decoder as is shown in Fig. 6. The encoder part is composed of a convolution subsampling layer and several conformer blocks. The role of conformer blocks is similar as that of transformer. A conformer block is composed of four modules: a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module in the end. The decoder part in conformer is also the same as in transformer.

In reference [6], conformer achieve 3 to 10% relative improvements over traditional transformer. The number of conformer blocks in encoder and decoder of our system is 12 and 6. The encoder and decoder dimension are both 2048. The attention layer contains 4 heads and 256 units per head. To complement the system fusion, we also use the full data of Babel to finetune the pre-training model introduced above. In addition, the performance of end to end model is degenerated when it comes to long sequences [15]. To solve this, we restrict the length of wave segments by VAD to be less than 15s.

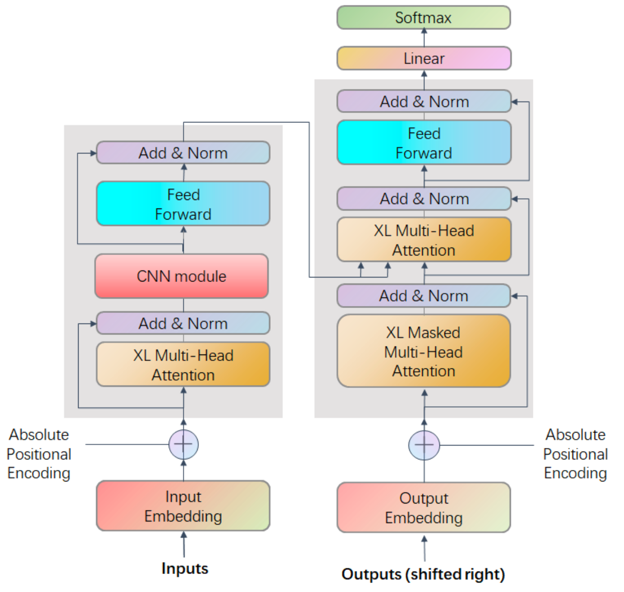

Figure 6 end to end structure of conformer model

B. Acoustic model for hybrid system

For acoustic model, we have accumulated additional training data for some languages. Unfortunately, most of the OpenASR DEV data is recorded in telephone channel with a sampling rate of 8kHz and we don't have any data that either matched for the target PSTN telephony condition or with the accent. Most of the extra data at our hands are wideband 16kHz speeches, in order to utilize these data, we first train a hybrid bandwidth acoustic model [16] as illustrated in Fig.7, resulting in a feature extractor for 16kHz. During inference stage, all the 8kh speeches should be up-sampled.

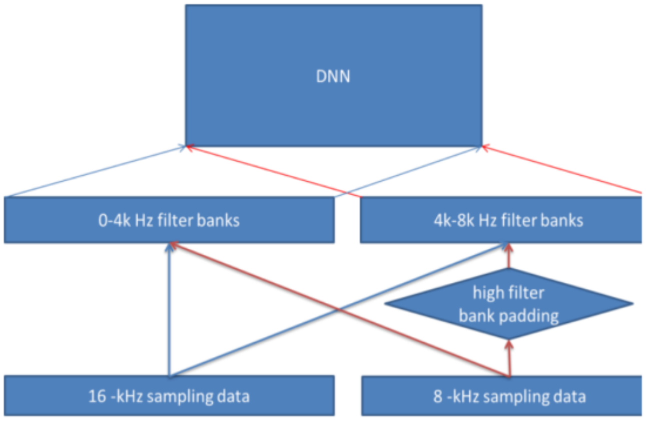

Figure 7 Multi-band speech models for 8-16kHz hybrid recognition

For Cantonese, 4000h wideband dictated speeches from Speech Ocean Inc, Datatang Inc, Aishell Inc and Huiting Tech Inc. [17,18] and 140h narrow band (8kHz) speeches from IARPA Babel are applied [19]. For Mongolian, 10h wide band speeches from Mozilla [20] and 50h narrow band speech from IARPA Babel are available. However, we believe that Inner Mongolian speech in China is also beneficial for Mongolian speech recognition, as a result, an extra 1000h wideband dictated speech from Speech Ocean [21] and 100h dictated speech from M2ASR project [22] are incorporated. Details are illustrated in Tab 4. For Kazakh, 1000h wide band speeches from Speech Ocean [21] and 50h narrow band speech from IARPA Babel are available.

TABLE4 EXTRA DATASETS FOR CANTONESE AND MONGOLIAN

	Narrowband(8khz)		Wideband(16khz)	
Language	**Data source**	**Duration**	**Data source**	**Duration**
Cantonese	IARPA Babel	~140h	Speech Ocean	~1000h
		~1000h	Huiting Tech	~1000h
			Datatang	~1000h
			Aishell	~1000h
Mongolian	IARPA Babel	~50h	Mozilla	10h
Inner Mongolian			Speech Ocean	~1000h
Kazakh	IARPA Babel	~50h	Speech Ocean	~1000h

All the acoustic model is trained with lexicon from IARPA Babel program, after which weight transfer is used in transfer learning to replace the output model layer. Finally, up-sampled OpenASR training data for the target language is used for fine-tuning. In addition, babel data is in parallel used to train acoustic model that supports 8kHz sampling rate, and mixed wideband e2e model is also trained and tuned towards same OpenASR data. The whole procedures for the main three streams of system are illustrated in Fig. 8.

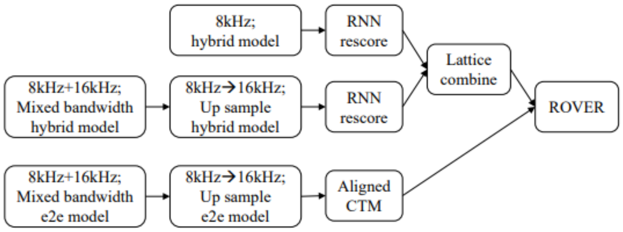
8kHz;
hybrid model
RNN rescore
Lattice combine
8kHz+16kHz; Mixed bandwidth hybrid model
8kHz→16kHz; Up sample hybrid model
RNN rescore
ROVER
8kHz+16kHz; Mixed bandwidth e2e model
8kHz→16kHz; Up sample e2e model
Aligned CTM

Figure 8 multi bandwidth hybrid & 16kHz end to end system fusions

C. *Language optimization*

Cantonese is a language within the Chinese language family. Since the Cantonese vernacular text data is scarce and irregular, we obtain certain additional text data through web crawling, and use regular methods to correct common errors in the text data, such as abbreviations and typos.

Text segmentation is needed to word sequence to calculate the word error rate in the evaluation. For e2e system, characters are used as the modeling unit rather than word. To solve this, we use all the crawled and Babel text data to train a text segmentation model through Cantonese BERT pre-training.

D. *System fusion*

For better performance, we build several different systems for system fusion as in Tab 5. All the above acoustic models in constrained system are trained using both 8kHz and hybrid band with data, which are fused with lattice combinations. [23]. For unconstrained conditions, an extra end-to-end system as mentioned above is built. Since that there is no lattice for end-to-end systems, CTM-level fusion is conducted for traditional hybrid systems and end-to-end systems through ROVER [14] as in Fig 6. The Cantonese results of all the fused systems in DEV set are as below, it can be observed that hybrid and end to end ASR system are in compensation with each other. Compared to using only e2e systems for fusion, the WER can be reduced from 0.316 to 0.313.

TABLE 5 UNCONSTRAINED SYSTEM FUSION RESULTS FOR CANTONESE (DEV SET, OUR COMPUTATION)

System	WER
Hybrid System Fusion (lattice fusion)	0.373
End-to-end System	0.324
End-to-end Pretrain System	0.326
Rover Fusion(e2e)	0.316
Rover Fusion (all CTMs)	0.313

E. *Results on eval set*

Our systems' performance of the unconstrained condition on the evaluation set is shown in Tab.6, which are released by NIST OpenASR scoring server. We notice that the results computed by dashboard in NIST OpenASR scoring server is much better than our scoring results in TABLE 5 (WER from 0.313->0.286), we realize that it is because we count language miscues, pauses, and other non-verbal speech as errors. By removing these constrains, we obtain similar results as in the dashboard.

TABLE 6 WER OF ASR SYSTEMS ON EVAL SET (UNCONSTRAINED)

	WER on DEV	WER on EVAL		
	Unconstrain	**Constrain**	**Constrain-plus**	**Unconstrain**
Cantonese	**0.286**	0.402	**0.337**	**0.277**
Mongolian	**0.285**	0.416	**0.378**	**0.342**
Kazakh	**0.312**		**0.427**	**0.394**

It can be observed that by using extra data, an absolute 12-13% WER reduction can be achieved such as Cantonese and the CER is even much lower than 0.23. In practice, speech recognition accuracy for Sino-Tibetan languages relies much more on CER rather than WER, WER for these types of languages is largely dominated by word segmentation error, which may incur a biased result.

V. HARDWARE AND TIME DESCRIPTION

The hardware of our proposed system is shown in Tab.7. As for the required time, the elapsed wall clock time is approximately 2 hours for one system of each language in constrained-plus condition and 20 hours in unconstrained condition. The corresponding total CPU time is about 10 hours, and the total GPU time is 2 hours or so.

TABLE 7 HARDWARE AND TIME DESCRIPTION

OS	CentOS 7.4 64-bit
CPU number	24
CPU description	112,Intel(R) Xeon(R) CPU E5-4650 v4 @ 2.20GHz
GPU number	8
GPU description	Tesla V100 SMX2 16GB * 40
RAM	256GB
RAM per CPU	128GB
Disk storage	About 3TB

REFERENCES

[1] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na,Y.Wang, and S. Khudanpur, "Purely sequence-trained neural networksfor ASR based on lattice-free MMI," in Interspeech. San Francisco, CA,USA: ISCA, Sep 2016, pp. 2751–2755.

[2] D. Povey, H. Hadian, P. Ghahremani et al., "A time-restricted self attention layer for ASR," in proc. ICASSP. Calgary, AB, Canada: IEEE,Apr. 2018, pp. 5874–5878.

[3] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deepneural networks." in Interspeech, 2018, pp. 3743–3747.

[4] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentationfor speech recognition," in INTERSPEECH. Dresden, Germany: ISCA,Sep 2015, pp. 3586–3589.

[5] Zhao, Jing, et al. "The TNT Team System Descriptions of Cantonese and Mongolian for IARPA OpenASR20," Proc. Interspeech 2021 (2021): 4344-4348.

[6] Gulati A, Qin J, Chiu C C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition[J].arXiv preprint arXiv:2005.08100, 2020.

[7] Baevski, Alexei and Zhou, et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv preprint arXiv:2006.11477.

[8] Zhang, Yu and Qin, James and Park, and et al. Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. arXiv preprint arXiv:2010.10504.

[9] Xiong Cai, Zhiyong Wu, Kuo Zhong,and et al. Unsupervised Cross-lingual Representation Learning for Speech Recognition. arXiv:2012.11174.

[10] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta ,et al. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks. arXiv:2004.10964

[11] https://huggingface.co/uer/albert-base-chinese-cluecorpussmall

[12] S. Kim, T. Hori, and S. Watanabe, “Joint ctc-attention based end-to-end speech recognition using multi-task learning, ” in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017, pp. 4835‑4839.

[13] Binbin Zhang, Di Wu, Zhuoyuan Yao,and et al. Unified Streaming and Non-streaming Two-pass End-to-end Model for Speech Recognition. arXiv:2012.05481.

[14] J. G. Fiscus, "A post-processing system to yield reduced word error rates:Recognizer output voting error reduction (rover)," in 1997 IEEEWorkshop on Automatic Speech Recognition and Understanding Proceedings. IEEE, 1997, pp. 347–354.

[15] Zhou P, Fan R, Chen W, et al. Improving Generalization of Transformerfor Speech Recognition with Parallel Schedule Sampling and RelativePositional Embedding[J]. arXiv preprint arXiv:1911.00203, 2019.

[16] Jinyu Li, Dong Yu, Jui-Ting Huang, Tifan Gong. Improving widebandspeech recognition using mixed-bandwidth training data in CD-DNN-HMM, IEEE Workshop on SLT, 2012

[17] http://www.speechocean.com/datacenter/details/709.html

[18] http://www.huitingtech.com/en/dataInfo.action?id=1005

[19] https://www.iarpa.gov/index.php/research-programs/babel

[20] https://pontoon.mozilla.org/mn/common-voice/project-info/

[21] http://www.speechocean.com

[22] Dong wang, et al. M2ASR: Ambitions and first year progress. O-COCOSDA. 2017

[23] Xu, H., Povey, D., Mangu, L., & Zhu, J. (2010, March). An improvedconsensus-like method for Minimum Bayes Risk decoding and lattice combination. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 4938-4941). IEEE.

[24] Wang, Zhichao, et al. "WNARS: WFST based Non-autoregressive Streaming End-to-End Speech Recognition." arXiv preprint arXiv:2104.03587 (2021).