WENETSPEECH: A 10000+ HOURS MULTI-DOMAIN MANDARIN CORPUS FOR SPEECH RECOGNITION

Binbin Zhang[1,2,], Hang Lv[1,*], Pengcheng Guo[1], Qijie Shao[1], Chao Yang[2], Lei Xie[1✉], Xin Xu[3], Hui Bu[3], Xiaoyu Chen[2], Chenchen Zeng[2], Di Wu[2], Zhendong Peng[2]*

[1] Audio, Speech and Language Processing Group (ASLP@NPU),
School of Computer Science, Northwestern Polytechnical University
[2]Mobvoi Inc. [3]Beijing Shell Shell Technology Co., Ltd.
hanglv@nwpu-aslp.org,lxie@nwpu.edu.cn

ABSTRACT

In this paper, we present *WenetSpeech*, a multi-domain Mandarin corpus consisting of 10000+ hours high-quality labeled speech, 2400+ hours weakly labeled speech, and about 10000 hours unlabeled speech, with 22400+ hours in total. We collect the data from YouTube and Podcast, which covers a variety of speaking styles, scenarios, domains, topics and noisy conditions. An optical character recognition (OCR) method is introduced to generate the audio/text segmentation candidates for the YouTube data on the corresponding video subtitles, while a high-quality ASR transcription system is used to generate audio/text pair candidates for the Podcast data. Then we propose a novel end-to-end label error detection approach to further validate and filter the candidates. We also provide three manually labelled high-quality test sets along with WenetSpeech for evaluation – *Dev* for cross-validation purpose in training, *Test_Net*, collected from Internet for *matched* test, and *Test_Meeting*, recorded from real meetings for more challenging *mismatched* test. Baseline systems trained with WenetSpeech are provided for three popular speech recognition toolkits, namely Kaldi, ESPnet, and WeNet, and recognition results on the three test sets are also provided as benchmarks. To the best of our knowledge, WenetSpeech is the current largest open-source Mandarin speech corpus with transcriptions, which benefits research on production-level speech recognition.

Index Terms— automatic speech recognition, corpus, multi-domain

1. INTRODUCTION

In the past decade, the performance of automatic speech recognition (ASR) systems have been significantly improved. On the one hand, the development of neural networks has increased the capacity of models, pushing the dominant framework from the hybrid hidden Markov models [1, 2] to end-to-end models, like CTC [3, 4], RNN-T [5, 6, 7, 8], and encoder-decoder based models [9, 10, 11, 12, 13]. To simply implement such advanced models and obtain state-of-the-art reproducible results, researchers also release several open source toolkits, including Kaldi [14], Sphinx [15], Fariseq [16], ESPnet [17], and recently WeNet [18], etc. On the other hand, self-supervised speech representation learning methods are proposed to make better use of a large amount untranscribed data, such as wav2vec [19], wav2vec 2.0 [20], Hubert [21], and wav2vec-U [22], etc. In addition to these algorithm-level efforts, the development of open source corpora is also crucial to the research community, especially for academia or small-scale research groups.

Most of the current open source speech corpora remain small size and lack of domain diversities. For example, the commonly used English speech corpus - Librispeech [23], which includes about 1000 hours reading speech from audiobook, currently has a word error rate (WER) of 1.9% on its test-clean benchmark. Recently, a large-scale multi-domain English corpus is released, which is named GigaSpeech [24]. It consists of 10000 hours of high quality transcribed speech for supervised ASR training and 40000 hours audio in total for semi-supervised or unsupervised training, contributing to the research community for developing more generalized ASR systems. Comparing with English corpora, the largest open source Mandarin speech data is AIShell-2 [25], including 1000 hours speech recorded in a quiet environment and having a state-of-the-art character error rate of 5.35%. It is too simple to do further research and ASR systems developed based on it may be susceptible to performance degradation in the complex real-world scenarios. In addition, current open source Mandarin corpora are also unable to train a well generalized pre-trained model, since both the Wav2vec 2.0 large model [20] and the XLSR-53 [26] use more than 50000 hours English speech data.

In this work, we release WenetSpeech, a large multi-domain Mandarin corpus. "We" means connection and sharing, while "net" means all of the data are collected from the Internet which is repository for diversity. The key features of WenetSpeech include:

- **Large scale**. 10000+ hours labeled data, 2400+ hours weakly labeled data, and about 10000 hours unlabeled data are provided, resulting in 22400+ hours audio in total.
- **Diverse**. The data are collected from multiple speaking styles, scenarios, domains, topics, and noisy conditions.
- **Extensible**. An extensible metadata is designed to extend the data in the future.

To the best of our knowledge, WenetSpeech is the current largest open source Mandarin speech corpus with domain diversity to satisfy various speech recognition tasks. In Section 2, we first introduce the construction procedure of WenetSpeech with a reliable pipeline to obtain high-quality transcriptions, including OCR-based caption recognition on Youtube video, ASR-based automatic transcription on Podcast audio, as well as a new end-to-end label error detection approach. The corpus composition is described in Section 3 and baseline benchmarks built on Kaldi, ESPnet and WeNet toolkits are introduced in Section 4. We believe our corpus will bring benefit to research community for developing more generalized ASR systems.

* Co-first authors, equal contribution. ✉ Corresponding author.

2. CREATION PIPELINE

In this section, we introduce the detailed creation pipeline of our WenetSpeech corpus, including original audio collection, audio/text segmentation candidate generation, and candidate calibration.

2.1. Stage 1: Audio Collection

In the beginning, we manually define the domains into 10 categories, which including *audiobook, commentary, documentary, drama, interview, reading, talk, variety, and others*. Then, we collected and tagged the audio files from YouTube and Podcast playlists according to our selected categories. Especially, for the YouTube data, the videos are also downloaded for preparing the audio/text segmentation candidates with our OCR system. For Podcast, since the manually transcribed Mandarin data is limited, we only considered the category information and prepared to transcribed it by a high-quality ASR system.

2.2. Stage 2: Candidates Generation

In this part, we introduce the specific pipelines to obtain the audio/text segmentation candidates from YouTube data by an OCR system and Podcast data by a high-quality ASR system. At last, text normalization[1] technique is applied to all the candidates.

2.2.1. YouTube OCR

As Figure 1 shown, an OCR-based pipeline is applied for generating candidates from embedded subtitles on YouTube videos.

1. *Text Detection*: apply CTPN [27] based text detection on each frame image in the video.
2. *Subtitle Validation*: mark frame t as the *start point* of a specific *subtitle phrase*, when a phrase of text is detected at the bottom of the screen for subtitles at frame t.
3. *Subtitle Change Detection*: compute the structural similarity (SSIM) of subtitle area frame by frame until a change is detected at frame $t+n$. Then, the frame $t+n-1$ is marked as the the *end point* of this *subtitle phrase*.
4. *Text Recognition*: a CRNN-CTC [28] based text recognition approach is used to recognize the detected subtitle area.
5. *Audio/Text Pair*: prepare each audio/text pair segmentation candidate with the corresponding (*start point, end point, subtitle phrase*) tuple.

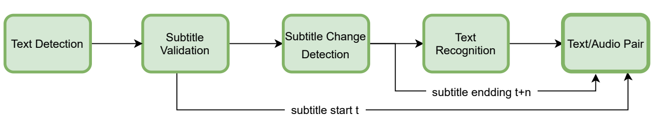

Fig. 1. OCR based YouTube data collection pipeline

To verify whether the proposed OCR-based pipeline is reliable, we sampled various subtitles and manually annotated them as benchmarks for testing the *Text Recognition* module. Finally, we obtain 98% accuracy on the test set and confirm the reliability of our OCR-based pipeline for WenetSpeech corpus.

Figure 2 shows 4 typical examples of our OCR-based system. The results of the *Text Detection* module are marked with boxes. If the box is in a reasonable subtitle area, it will be marked as red, and then, the subtitle box is further processed by the *Text Recognition* module. At last, the recognition result is shown above each subtitle box.

In addition, we find that some annotators prefer to split a long subtitle phrase into many pieces for a video, so that the problem of audio and subtitle asynchronous is introduced. This leads an inaccurate subtitle boundary detection problem to our OCR system. To alleviate it, we merge the consecutive video phrases until the audio is over 8 seconds.

(a) Audiobook (b) Game commentary

(c) Drama (d) Lecture

Fig. 2. Example outputs of the OCR pipeline

2.2.2. Podcast Transcription

We use a third-party commercial ASR transcription system to transcribe all the Podcast data. The transcription system is one of the best system on the public benchmark platform[2], and more than 95% accuracy rates have been achieved in most of the testing scenarios, including news, reading, talk show, conversation, education, games, TV, drama, and so on.

The transcription system first segments the original audio into short segments, and then the audio/text pair segmentation candidates are generated by speech recognition.

2.3. Stage 3: Candidates Validation

Although the used OCR system and transcription system are high-quality enough, the errors from candidate generation, such as subtitle annotation error, timestamp inaccuracy, OCR mistake, transcription word error, and text normalization error, are still unavoidable. To further improve the quality of our WenetSpeech corpus, we apply the following validation approach to classify the candidates according to their confidences and filter out the extremely bad candidates.

2.3.1. Force Alignment Graph

Here, we propose an novel CTC-based end-to-end force alignment approach to detect the transcription error. The transcription is first segmented by the model unit of CTC, and then a *force alignment graph* (L) is built for each candidate as shown in Figure 3. The key features of the alignment graph are:

1. The oracle transcription alignment path is included.
2. Arbitrary deletion operation at any position is allowed through tag $\langle \text{del} \rangle$ with penalty p_1.
3. Arbitrary insertion or substitution at any position is allowed. From each reasonable position, a start tag $\langle \text{is} \rangle$ and a end tag $\langle /\text{is} \rangle$ are connected to a global *filler* state. On this *filler* state, each CTC modeling unit has a corresponding self-loop arc with penalty p_2, which is presented by tag $\langle \text{gbg} \rangle$.

This makes it possible to capture the error between the audio and the corresponding transcription through decoding technique.

[1]https://github.com/speechio/chinese_text_normalization

[2]https://github.com/SpeechColab/Leaderboard

Compared with traditional hybrid validation approach which is used in Librispeech, our proposed approach is a pure end-to-end approach. There is no need for HMM typologies, lexicon, language model components, or careful design of the filler state. So the proposed novel approach simplifies the whole pipeline a lot. The force alignment graph is implemented by the WeNet toolkit, and it is publicly available [3].

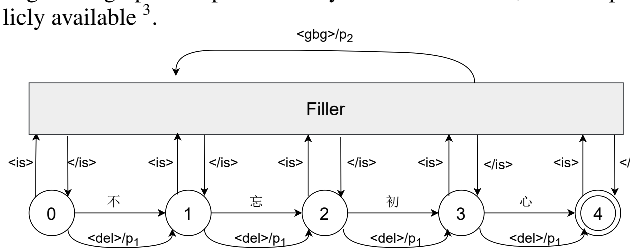

Fig. 3. An example force alignment graph L of "不忘初心"

2.3.2. Label Error Detection

After defining the force alignment graph L, it is further composed with the CTC topology graph T [29] to build the final *force decoding graph* $F = T \circ L$ for label error detection and validation.

In addition, we assign confidence for each candidate by it's reference (ref) and the force decoding hypothesis (hyp). The confidence c is computed as

$$c = 1 - \frac{\text{EditDistance}(ref, hyp)}{\max(\text{len}(ref), \text{len}(hyp))}.$$

With the confidence, we classify the audio/text segmentation candidates of our WenetSpeech corpus into *Strong Label* and *Weak Label* sets and even filter out the extremely ones to *Others* set. Figure 4 shows two real examples that we find by applying label error detection. In *Case 1*, human subtitle error was successfully detected. In *Case 2*, OCR error was successfully detected.

Please note the model used in the transcription system and force alignment system are developed and trained with different model methods and data. p_1 is set to 2.3 and p_2 is set to 4.6 in our pipeline.

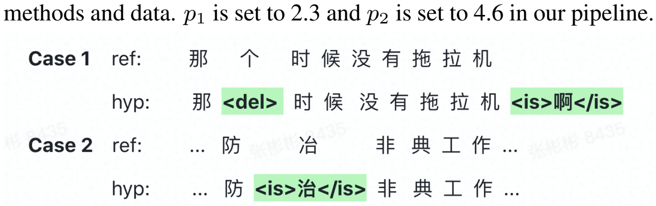

Fig. 4. Examples of label error detection

3. THE WENETSPEECH CORPUS

In this section, we describe the metadata, audio format, partition by confidence, data diversity, and training and evaluation set design of our WenetSpeech corpus. Instructions and scripts are available at WenetSpeech GitHub repo[4].

3.1. Metadata and Audio Format

We save all the metadata information to a single JSON file. Local path, original public URL, domain tags, md5, and segments are provided for each audio. And timestamp, label, confidence, subset information are provided for each segment. The design is extensible, and we are planning to add more diverse data in the future.

The original audio files are downloaded and converted to 16k sampling rate, single-channel, and 16-bit signed-integer format. Then Opus compression is applied at an output bit rate of 32 kbps to reduce the size of the WenetSpeech corpus.

Table 1. WenetSpeech partition

Set	Confidence	Hours
Strong Label	[0.95, 1.00]	10005
Weak Label	[0.60, 0.95)	2478
Others	/	9952
Total(hrs)	/	22435

Table 2. Training data in different domains with duration (hrs)

Domain	Youtube	Podcast	Total
audiobook	0	250.9	250.9
commentary	112.6	135.7	248.3
documentary	386.7	90.5	477.2
drama	4338.2	0	4338.2
interview	324.2	614	938.2
news	0	868	868
reading	0	1110.2	1110.2
talk	204	90.7	294.7
variety	603.3	224.5	827.8
others	144	507.5	651.5
Total	6113	3892	10005

3.2. Size and Confidence

We assign confidence for each valid segment which measures the label quality, where the confidence in defined in Section 2.3.2. As shown in Table 1, we select 10005 hours *Strong Label* data, whose confidence is greater than 0.95, as the supervised training data. The 2478 hours *Weak Label* data, whose confidence is between 0.60 and 0.95, is reserved in our metadata for semi-supervised or other usage. At last, *Others* represent all invalid data (i.e. the confidence of data is less than 0.6 or unrecognized) for speech recognition task. In summary, WenetSpeech has 22435 hours of raw audio.

3.3. Training Data Diversity and Subsets

We tag all the training data with its source and domain. All of the training data is from *Youtube* and *Podcast*. As shown in Table 2, we classify the data into 10 groups according to its category. Please note about 4k hours of data is from drama, which is a special domain with a wide range of themes, topics and scenarios, which may cover any kind of other categories.

As shown in Table 3, we provide 3 training subsets, namely *S*, *M* and *L* for building ASR systems on different data scales. Subsets *S* and *M* are sampled from all the training data which have the oracle confidence 1.0.

3.4. Evaluation Sets

We will release the following evaluation datasets associated with WenetSpeech, and the major information is summarized in Table 4.

1. Dev, is specifically designed dataset for some speech tools which require cross-validation in training.
2. Test_Net, is a *matched* test set from the internet. Compared with the training data, it also covers several popular and difficult domains like game commentary, live commerce, etc.
3. Test_Meeting, is a *mismatched* test set since it is a far-field, conversational, spontaneous, and meeting speech dataset. It is sampled from 197 real meetings in a variety of rooms. Its

[3]https:github.com/wenet-e2e/wenet/blob/main/runtime/core/bin/label_checker_main.cc

[4]https://github.com/wenet-e2e/WenetSpeech

Table 3. The training data subsets

Training Subsets	Confidence	Hours
L	[0.95, 1.0]	10005
M	1.0	1000
S	1.0	100

Table 4. The WenetSpeech evaluation sets

Evaluation Sets	Hours	Source
Dev	20	Internet
Test_Net	23	Internet
Test_Meeting	15	Real meeting

topics cover education, real estate, finance, house and home, technology, interview and so on.

The three evaluation sets are carefully checked by professional annotators to ensure the transcription quality.

4. EXPERIMENTS

In this section, we introduce the baseline systems and experimental results on three popular speech recognition toolkits, Kaldi [14], ESPnet [17] and WeNet [18].

4.1. Kaldi Benchmark[5]

The Kaldi baseline implements a classical chain model [30] using various amounts of WeNetSpeech data (i.e. S, M, L). We choose the open source vocabulary, BigCiDan[6], as our lexicon. And we segment our transcriptions with the open source word segmentation toolkit, jieba [31]. First, we train a GMM-HMM model to obtain the training alignments. Then, we train a chain model, which stacks 6 convolutional neural network (CNN) layers, 9 factored time-delay neural network (TDNN-F) layers [32], 1 time-restricted attention layer [33] ($H = 30$), 2 projected long short-term memory (LSTMP) with TDNN-F blocks. We feed the 40 dimensional filter-bank (FBank) features and 100 dimensional i-vector features as the input. In order to be consistent with the other systems, we only use SpecAugment [34] technique and abandon the speed/volume perturbation techniques. The chain model is trained by LF-MMI criterion with cross-entropy loss (10 epochs for subset *S*, and 4 epochs for subset *M* and *L*). A 3-gram language model (LM) is used for decoding and generating the lattice. Finally, a recurrent neural network LM (RNNLM) is further adopted to rescore the lattices. The 3-gram and RNNLM are both trained on all the WenetSpeech transcriptions.

4.2. ESPnet Benchmark[7]

The ESPnet baseline employs a Conformer model [13, 35] which is designed to capture the global context with the multi-head self-attention module and learn the local correlations synchronously with the convolution module. Our Conformer model consists of a 12-block Conformer encoder ($d^{\text{ff}} = 2048, H = 8, d^{\text{att}} = 512, \text{CNN}_{\text{kernel}} = 15$) and a 6-block Transformer [36] decoder ($d^{\text{ff}} = 2048, H = 8$). A set of 5535 Mandarin characters and 26 English letters is used as the modeling units. The objective is a logarithmic linear combination of the CTC ($\lambda = 0.3$) and attention objectives. Label smoothing is applied to the attention objective. During data preparation, we generate 80-dimensional FBank features with a 32ms window and a 8ms frame shift. SpecAugment with 2 frequency masks ($F = 30$) and 2 time masks ($T = 40$), and global CMVN technique are used as data pre-processing. During training, we choose the Adam optimizer with the maximum learning rate of 0.0015. The Noam learning rate scheduler with 30k warm-up steps is used. The model was trained with dynamic batching skill for 30 epochs. At last, the last 10 best checkpoints were averaged to be the final model. For decoding, the ESPNet system follows the joint CTC/Attention beam search strategy [37].

4.3. WeNet Benchmark[8]

The WeNet baseline implements a U2 model [18], which unifies streaming and non-streaming end-to-end (E2E) speech recognition in a single model. The basic setup of our WeNet model is same as the ESPnet model except the following minor points: 1) We prepare 80-dimensional FBank features with a 25ms window and a 10ms frame shift. SpecAugment with 2 frequency masks ($F = 30$) and 3 time masks ($T = 50$) and global CMVN technique are applied on the top of our features. 2) The max trainable epoch is 25. Models of the last 10 epochs were averaged to be the final model. The key difference between WeNet and ESPNet is different decoding strategies. Specifically, different from ESPNet's auto-regressive decoding strategy, Wenet generates the N-Best hypotheses by the CTC branch and rescores them by the attention branch.

Table 5. Results (MER%) on different test sets for baseline systems trained using WenetSpeech training subset L

Toolkit	Dev	Test_Net	Test_Meeting	AIShell-1
Kaldi	9.07	12.83	24.72	5.41
ESPNet	9.70	8.90	15.90	3.90
WeNet	8.88	9.70	15.59	4.61

Table 6. Kaldi baseline results (MER%) for different WenetSpeech training subsets

SubSet	Dev	Test_Net	Test_Meeting	AIShell-1
L	9.07	12.83	24.72	5.41
M	9.81	14.19	28.22	5.93
S	11.70	17.47	37.27	7.66

4.4. Experimental Results

We must announce that the results listed here are purely for the purpose of providing a baseline system for each toolkit. They might not reflect the state-of-the-art performance of each toolkit.

In Table 5, we report the experimental results in Mixture Error Rate (MER) [38], which considers Mandarin characters and English words as the tokens in the edit distance calculation, on three designed test sets and one well-known, publicly available test set (i.e. AIShell-1 [39] test) with Kaldi, ESPNet and WeNet toolkits respectively. The good performance on AIShell-1 reflects the diversity and reliability of our WenetSpeech corpus. And the results on our designed test sets reflect they are quite challenging. In Table 6, we provide the Kaldi baseline results for difference scale WenetSpeech subsets. As the growth of the data amount, the performance goes up steadily.

5. ACKNOWLEDGEMENTS

We thank Jiayu Du and Guoguo Chen for their suggestions on this work. We thank Xi'an Future AI Innovation Center for providing hosting service for *WenetSpeech*. We also thank MindSpore for the support of this work, which is a new deep learning computing framwork[9]. Our gratitude goes to Lianhui Zhang and Yu Mao for collecting some of the YouTube data.

[5] https://github.com/wenet-e2e/WenetSpeech/tree/main/toolkits/kaldi

[6] https://github.com/speechio/BigCiDian

[7] https://github.com/wenet-e2e/WenetSpeech/tree/main/toolkits/espnet

[8] https://github.com/wenet-e2e/WenetSpeech/tree/main/toolkits/wenet

[9] https://www.mindspore.cn/

6. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, 2012.

[2] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[3] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. International Conference on Machine Learning (ICML)*, 2016, pp. 173–182.

[5] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.

[7] Xiong Wang, Zhuoyuan Yao, Xian Shi, and Lei Xie, "Cascade rnn-transducer: Syllable based streaming on-device mandarin speech recognition with a syllable-to-character converter," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 15–21.

[8] Xiong Wang, Sining Sun, Lei Xie, and Long Ma, "Efficient conformer with prob-sparse attention mechanism for end-to-endspeech recognition," in *Proc. ISCA Interspeech*. ISCA, 2021, pp. 4578–4582.

[9] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2015, pp. 577–585.

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4960–4964.

[11] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[12] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. ISCA Interspeech*. ISCA, 2020, pp. 5036–5040.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE Signal Processing Society, 2011.

[15] K-F Lee, H-W Hon, and Raj Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35–45, 1990.

[16] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019, pp. 48–53.

[17] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. ISCA Interspeech*. ISCA, 2018, pp. 2207–2211.

[18] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. ISCA Interspeech*. ISCA, 2021, pp. 4054—-4058.

[19] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. ISCA Interspeech*. ISCA, 2019, pp. 3465—-3469.

[20] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[21] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6533–6537.

[22] Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli, "Unsupervised speech recognition," *arXiv preprint arXiv:2105.11084*, 2021.

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan, "Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[25] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "Aishell-2: Transforming mandarin ASR research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[26] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[27] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao, "Detecting text in natural image with connectionist text proposal network," in *Porc. European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 56–72.

[28] Baoguang Shi, Xiang Bai, and Cong Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[29] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[30] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. ISCA Interspeech*. ISCA, 2016, pp. 2751–2755.

[31] J Sun, "Jieba chinese word segmentation tool," 2012.

[32] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Proc. ISCA Interspeech*. ISCA, 2018, pp. 3743–3747.

[33] Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li, and Sanjeev Khudanpur, "A time-restricted self-attention layer for ASR," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5874–5878.

[34] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. ISCA Interspeech*. ISCA, 2019, pp. 2613–2617.

[35] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, Jing Shi, Shinji Watanabe, Kun Wei, Wangyou Zhang, and Yuekai Zhang, "Recent developments on ESPnet toolkit boosted by conformer," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[37] Takaaki Hori, Shinji Watanabe, and John R Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 518–529.

[38] Xian Shi, Qiangze Feng, and Lei Xie, "The ASRU 2019 mandarin-english code-switching speech recognition challenge: open datasets, tracks, methods and results," *arXiv preprint arXiv:2007.05916*, 2020.

[39] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.