

# THE SLT 2021 CHILDREN SPEECH RECOGNITION CHALLENGE: OPEN DATASETS, RULES AND BASELINES

Fan Yu<sup>1\*</sup>, Zhuoyuan Yao<sup>1\*</sup>, Xiong Wang<sup>1</sup>, Keyu An<sup>2</sup>, Lei Xie<sup>1</sup>, Zhijian Ou<sup>2</sup>,  
Bo Liu<sup>3</sup>, Xiulin Li<sup>3</sup>, Guanqiong Miao<sup>3</sup>

<sup>1</sup>Audio, Speech and Language Processing Group,  
School of Computer Science, Northwestern Polytechnical University, Xi'an, China

<sup>2</sup>Speech Processing and Machine Intelligence Lab,  
Tsinghua University, Beijing, China

<sup>3</sup>Databaker Technology, Beijing, China

## ABSTRACT

Automatic speech recognition (ASR) has been significantly advanced with the use of deep learning and big data. However improving robustness, including achieving equally good performance on diverse speakers and accents, is still a challenging problem. In particular, the performance of children speech recognition (CSR) still lags behind due to 1) the speech and language characteristics of children's voice are substantially different from those of adults and 2) sizable open dataset for children speech is still not available in the research community. To address these problems, we launch the Children Speech Recognition Challenge (CSRC), as a flagship satellite event of IEEE SLT 2021 workshop. The challenge will release about 400 hours of Mandarin speech data for registered teams and set up two challenge tracks and provide a common testbed to benchmark the CSR performance. In this paper, we introduce the datasets, rules, evaluation method as well as baselines.

**Index Terms**— Automatic speech recognition, children speech recognition, deep learning, audio, datasets

## 1. INTRODUCTION

Automatic speech recognition (ASR) technology has been advanced with the recent help from deep learning (DL) [1]. A robust speech recognition system is indispensable to a large number of applications in human-computer interaction, on-line education and audio content analysis. In a broad view, *robustness* means the system has equally good performance for adverse acoustic conditions, diverse speakers and accents, different speaking styles and many other factors. However, current DL-based ASR systems are data-hungry, which is easy to overfit with limited training data. In particular, the speech and language characteristics of children's voices are substantially different from those of adults and ASR for children is

still significantly less accurate than that of adults without the coverage of a huge amount of labeled children's speech during training [2]. Recognizing children's speech has substantial applications, such as computer-assisted language learning (CALL), speech-enabled smart toys, etc.

The difficulty of children's speech recognition lies in the differences in anatomy and language expression between children and adults. Anatomically, the vocal cords of children are shorter and lighter than those of adults. These lead to higher fundamental and resonant frequencies and greater spectral variability of children's speech [3, 4]. In terms of language expression, children's language skills are obviously worse than those of adults. In particular, children in younger age are unable to express accurately and inevitably have mispronunciations due to limited and evolving language knowledge. Differences also exist between children's and adult's pronunciation in prosody, form, syntax, pragmatics and many other aspects [5]. In addition, another bottleneck to the current development of children speech recognition (CSR) is the lack of sizable labelled children's speech data in the research community.

At the early stage of ASR development, researchers mainly considered the particularity of children's speech and focused on the effect of language models in children's speech recognition. Compared with the adult model, the language model of children's speech is more helpful to children's speech recognition, which showed that there are differences in language expression between children and adults [6]. Some researchers [7] also customized specific pronunciation dictionaries for children to improve CSR performance. Fine-tuning the adult language model using child speech data also gained improvements [8]. In recent years, with the development of deep learning and the maturity of ASR technology, deep neural network (DNN) provided a great improvement to the acoustic model which has been widely adopted by the research community. DNN can better approach the nonlinear function requested by speech modeling through training with

The first two authors contributed equally to this work.

**Table 2.** Challenge track setting.

	Track 1	Track 2
Acoustic model training data	Set A, C1 and C2	Set A, C1, C2 and external data appeared in openslr
Language model training data	The text transcription in Set A, C1 and C2	The text transcription in Set A, C1, C2 and external data appeared in openslr
Evaluation data	Over 10 hours of Children’s reading and conversational speech	

a large amount of data, thus surpassing the ASR system based on Gaussian mixture model (GMM). However, due to the lack of a sizable dataset on children’s speech, there are only a few works on DNN to study children’s speech recognition. According to [9, 10], training the ASR system based on hybrid DNN-HMM led to slightly improved performance compared with the traditional GMM-based system. There was another method which used a DNN to predict the frequency warping factors for vocal tract length normalization (VTLN) for children, which brought improvement on hybrid DNN-HMM CSR system [11]. Convolutional long short-term memory recurrent neural networks can make great progress on children speech recognition [12], and augmenting data by artificially adding noise also achieve stronger robustness. Recent studies [12, 13, 14, 15] combined adult speech with children speech for model training while some researchers [16] employed multi-task learning (MTL) which can improve performance for both adults and children. Transfer learning and MTL were also found useful for the CSR task [17, 18].

The children Speech Recognition Challenge (CSRC) is a flagship satellite event of IEEE SLT 2021 workshop<sup>1</sup>. The challenge aims at advancing CSR by releasing a sizable set of labeled children speech data and providing a common testbed to benchmark the CSR performance. Specifically, we will release 400 hours of Mandarin speech data for system building, including 340 hours of adult speech, 30 hours of children reading speech as well as 30 hours of children conversational speech. We particularly set up two challenge tracks: the first track limits the usage of data released by the challenge while the second track allows the participants to use extra constrained data. Performance ranking is based on the character error rate (CER) on the evaluation dataset which is composed of both children reading and conversational speech. Rules and three baselines (Chain model [19], Transformer [20] and CTC-CRF [21]) are also provided.

The rest of this paper is organized as follows, section 2 will illustrate the released datasets. Section 3 will describe the track setting and evaluation plan. The baseline systems are discussed in Section 4. The challenge rules and other logistics are described in Section 5.

**Table 1.** Basic information about the released dataset. A: adult speech training set; C1: children speech training set; C2: children conversation training set.

Dataset	A	C1	C2
Duration (hrs)	341.4	28.6	29.5
# Speakers	1999	927	54
Speaker age	18-60	7-11	4-11
Speaking style	Reading	Reading	Conversational
Format	16kHz, 16bit, single channel wav		

## 2. DATASETS

Our goal of releasing the open source datasets in this challenge is to ensure the common system building resources and fair evaluation for participants. To this end, we will release single-channel adult speech, child reading and conversational speech for training (totally 400 hours) during the system building stage and additional children’s reading and conversational speech for evaluation and ranking in the evaluation stage. The basic information of the training data to be released is summarized in Table 1.

## 3. TRACK SETTING AND EVALUATION

We set up two tracks as shown in Table 2 for participants to study children speech recognition with different limits on the scope of training data. In the first track, participants can only use the data provided by the challenge to train acoustic and language models. In addition, the second track allows the participants to use the data listed in *openslr*<sup>2</sup>. Note that the language model training in the second track allows the use of transcripts associated with the provided speech data and the external speech data in *openslr*.

There are some differences in the files that need to be submitted between the two tracks. For both tracks, participants need to provide detailed technical solutions and decoding results on the evaluation set, including but not limited to the model structure and algorithms used. Besides, for the sec-

<sup>1</sup><http://2021.ieeeslt.org/>

<sup>2</sup><http://www.openslr.org/>

ond track, participants need to provide a list of additional data used during training.

We use edit distance based CER to evaluate the model performance for both track 1 and 2. The character error rate is calculated as: the summed number of error characters in both the children reading evaluation set and the children conversational evaluation set divided by the total number of characters in the two evaluation sets, as shown in Eq. (1):

$$Score_{CER} = \frac{CE_{reading} + CE_{conversation}}{CN_{total}} \times 100\%, \quad (1)$$

where  $CE_{reading}$  is the number of character errors on the children reading evaluation set,  $CE_{conversation}$  is the number of character errors on the children conversational evaluation set and  $CN_{total}$  is the total number of characters in the evaluation set. As standard, insertion, deletion and substitution all account for the errors.

## 4. BASELINE SYSTEMS

### 4.1. Chain-model based hybrid system

As a baseline, we implement a chain model based hybrid system using Kaldi [22]. The chain model is a type of DNN-HMM model, but there are two differences between chain model and the common DNN model. Firstly, instead of a frame-level objective, chain model uses the log-probability of the phone sequences as the objective function, and computes numerator and denominator under the maximum mutual information (MMI) criterion. Secondly, as for the HMM topology, chain model has a special state accepting phoneme only once and another state accepting silence zero or more times. We briefly review the chain model in the following. Readers can refer to [19] for more details.

The chain model uses the MMI criterion as the objective function. Specifically, in the conventional speech recognition, we only optimize  $\sum_u P_\theta(X|Y)$  in the acoustic model, while  $X = (x_1, x_2, \dots, x_n)$  is the input feature sequence and  $Y = (y_1, y_2, \dots, y_m)$  is the prediction sequence. But in MMI we hope to optimize the whole formula, and further decompose the denominator with the full probability formula to Eq. (2) in which  $\theta_{MMI}$  is the objective model parameter.

$$\begin{aligned} \theta_{MMI} &= \underset{\theta}{\operatorname{argmax}} \sum_u \log P_\theta(Y|X) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_u \log \frac{P_\theta(X|Y)P(Y)}{P(X)} \\ &= \underset{\theta}{\operatorname{argmax}} \sum_u \log \frac{P_\theta(X|Y)P(Y)}{\sum_W P_\theta(X|Y)P(Y)} \end{aligned} \quad (2)$$

Moreover, the chain model uses a model topology different from the conventional HMM which is a 3-state left-to-right HMM that can be traversed in a minimum of 3 frames.

The topological structure of chain will output a label representing a phoneme for the first frame of a phoneme, while the output of other frames of the phoneme is similar to blank in the connectionist temporal classification (CTC) [23].

To create the denominator graph, we need a language model and an N-gram model is usually used. Tri-gram and 4-gram are popular choices.

### 4.2. Transformer system

Besides the conventional hybrid system, we also use a transformer-based end-to-end model as another baseline, which is based on an encoder-decoder structure with an attention mechanism. Transformer was firstly introduced in neural machine translation and achieved superior performance recently in speech recognition [24, 25, 26, 27]. As shown in Fig. 1, the encoder can be regarded as a feature extractor that maps the input feature sequence  $X = (x_1, x_2, \dots, x_n)$  to a higher representation sequence  $H = (h_1, h_2, \dots, h_n)$ . The decoder decodes  $H$  in the way of autoregressive and gets the prediction sequence  $Y = (y_1, y_2, \dots, y_m)$  step by step. The modeling unit of prediction sequence  $Y$  needs to be adjusted according to different ASR tasks [28, 29] in order to obtain a better recognition performance. The commonly used modeling units are phoneme, syllable, character, grapheme or wordpiece.

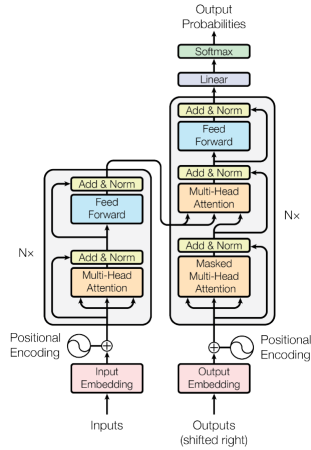


Fig. 1. Transformer Model [20]

The encoder and decoder of the transformer are composed of stacked self-attention and position-wise feed-forward layers. In the transformer model multi-head attention plays a significant role, which can obtain information from different representation subspace and each head could focus on a different subspace. Besides, the transformer does not contain the recurrent neural network (RNN) layer, so it cannot obtain any position information of the acoustic sequence. In order to enable the model to make rational use of the location information of the sequence, it is necessary to add the position

coding structure to obtain the location information of the corresponding symbols of the whole sequence [20].

### 4.3. CTC-CRF based system

The third baseline system is based on the recently proposed CTC-CRF [30]. Similar to CTC, CTC-CRF uses a hidden state sequence to obtain the alignment between the label sequence and input feature sequence, and the hidden state sequence is mapped to a unique label sequence by removing consecutive repetitive labels and blanks. The posterior of the hidden state sequence is defined by a conditional random field (CRF), and the potential function of the CRF is defined as the sum of node potential and edge potential. The definition of node potential is the same as CTC, and the edge potential is realized by an n-gram LM of labels. By incorporating the probability of the sentence into the potential, the conditional independence between the hidden states in CTC is naturally avoided.

### 4.4. Experiment Setup

We conduct experiments on the released dataset for a total of 400 hours of speech. We manually separate 1% of the data as a development set and the rest as a training set to benchmark the three baseline systems. We ensure there is no overlap of speakers between the training and development sets. The development set includes adult reading (AR), child reading (CR) and child conversation (CC) subsets.

In our baseline chain-model system, 71-dimensional mel-filterbanks features are used as the input of the acoustics model and frame length is 25 ms with a 10 ms shift. The acoustic model consists of a single convolutional layer with the kernel of 3 to down-sample the input speech feature, 19 hidden layers of a 256-dimensional Time Delay Neural Network (TDNN) and a 1280-dimensional fully connection layer with activation function of ReLU to obtain the information in time domain and map the speech features to high-dimensional representation. Finally, we use a fully connected layer with ReLU activation function and a log softmax layer to map the feature to the posterior probability of pdf-id. Meanwhile, we train a tri-gram language model using the allowed transcripts and out-of-vocabulary (OOV) words are mapped into  $< unk >$ . After combining the HMM model, the size of the final ‘HCLG’ graph is 384M.

We use the ESPnet [31] end-to-end speech processing toolkit to build our transformer baseline. The input features are the same as the chain model. We follow the standard configuration of the state-of-the-art ESPnet transformer which contains 12-layer encoder and 6-layer decoder, where the dimension of attention and feed-forward layer is set to 256 and 2048 respectively. In all attention sub-layers, 4 heads are used. The whole network is trained for 50 epochs and warm-up [20] is used for the first 25,000 iterations. We use 5140 commonly used Chinese characters as the modeling units.

We use CAT [21] to build our CTC-CRF baseline. The AM network is two blocks of VGG layers followed by a 6-layer BLSTM. The first VGG block has 3 input channels corresponding to spectral features, delta, and delta delta features. The BLSTM has 512 hidden units per direction and the total number of parameters is 37M. In training, a dropout probability of 0.5 is applied to the LSTM to prevent overfitting. A CTC loss with a weight 0.01 is combined with the CRF loss to help convergence. We use the open-source Aishell lexicon<sup>3</sup> as our basic lexicon, and then segment the training transcription, and add the OOV (out-of-vocabulary) words to the lexicon. The language model is a tri-gram word based language model, which is trained on the training transcription after segmentation. The size of decoding graph is 142M.

### 4.5. Results and Analysis

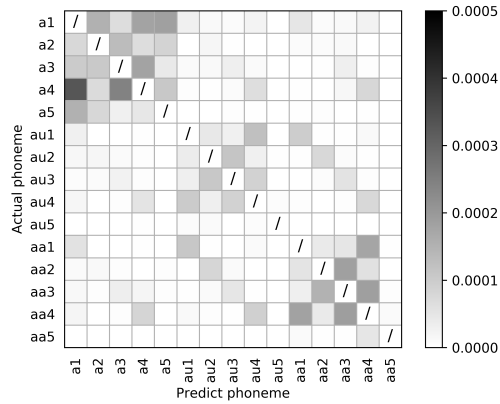
First of all, we mix all the children and adult data together for training. Table 3 reports the results of chain-based hybrid model on Kaldi, transformer-based end-to-end system on ESPnet and CTC-CRF-based system on CAT. According to Table 3, the CER on CC is significantly worse than that of the other two sets. This indicates that children’s conversational speech is more challenging to recognize. We further try simple model adaptation. Specifically, for the three systems, we merge children training data from C1, C2 and the same amount of adult data from Set A. Then we use the merged data to fine-tune the model B0, B1 and B2. As shown in Table 3, the CER of the chain model, the transformer model and the CTC-CRF model on the CC set decreases relatively by 2.5%, 5.4% and 0.9% respectively. Meanwhile, the evaluation criteria  $Score_{CER}$  decreases relatively by 2.3%, 3.1% and 1.3% respectively.

**Table 3.** Result comparison on development set in CER (%).

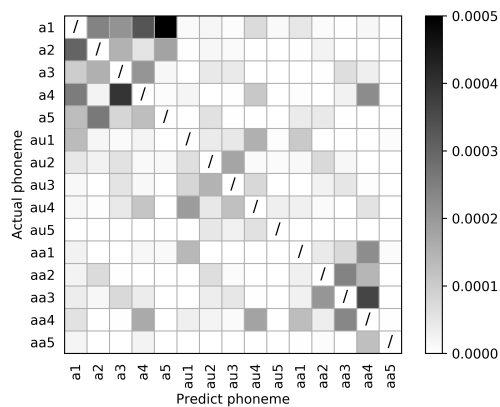
Exp ID	Model	CER (%)			
		AR	CR	CC	Score <sub>CER</sub>
B0	Chain model	11.5	12.2	36.4	26.6
B0A	+ fine-tune	12.0	12.1	35.5	26.0
B1	Transformer	12.5	10.9	35.5	25.5
B1A	+ fine-tune	15.0	11.8	33.6	24.7
B2	CTC-CRF	9.6	9.8	33.2	23.7
B2A	+ fine-tune	9.8	9.7	32.9	23.4

In order to further analyze the challenge of children’s speech compared with adult speech, we use all the training data to train a phoneme classifier. Fig. 2(a) shows the confusion matrix of some phonemes on the AR and CR development sets. It can be observed that compared with AR, phoneme discriminators are more likely to confuse similar phonemes on the CR set. This phenomenon is more obvious within  $\{a, aa, au\}$  which are similar vowel phonemes.

<sup>3</sup><http://www.openslr.org/resources/33/resource.aishell.tgz>



(a) AR set.



(b) CR set.

**Fig. 2.** Confusion matrix of some phonemes on two sets. Note that diagonal is not filled for better showing the phoneme confusion.

## 5. CHALLENGE RULES AND SCHEDULE

### 5.1. Rules

All participants should adhere to the following rules to be eligible for the challenge.

1. Data augmentation is allowed on the original training dataset, including, but not limited to, adding noise or reverberation, speed perturb and tone change.
2. The use of evaluation datasets in any form of non-compliance is strictly prohibited, including but not limited to use the evaluation dataset to fine-tune or train the model.
3. Multi-system fusion is not allowed.
4. If the CER of the two systems on the evaluation dataset are the same, the system with lower computation complexity will be judged as the superior one.

5. If forced alignment is used to obtain the frame-level classification label, the forced alignment model must be trained on the basis of the data allowed by the corresponding track.
6. Shallow fusion is allowed for the end-to-end approaches, e.g., LAS, RNNT and Transformer, but the training data of the shallow fusion language model can only come from the transcripts of the training dataset.
7. The right of final interpretation belongs to the organizer. In case of special circumstances, the organizer will coordinate the interpretation.

### 5.2. Timeline

- *August 14<sup>th</sup>, 2020* : Registration deadline, the due date for participants to join the Challenge.
- *August 21<sup>st</sup>, 2020* : Training data release.
- *September 30<sup>th</sup>, 2020* : Evaluation data release.
- *October 10<sup>th</sup>, 2020* : Final submission deadline.
- *October 25<sup>th</sup>, 2020* : Evaluation result and ranking release.
- *November 8<sup>th</sup>, 2020* : Evaluation data release.
- *January 19<sup>th</sup>–22<sup>nd</sup>, 2021* : SLT2021 main workshop and challenge workshop.

## 6. CONCLUSIONS

In this paper, we introduce the necessity of IEEE SLT 2021 children speech recognition challenge (CSRC) and describe the datasets, tracks, rules and baselines for the challenge. With the released sizable open data and common testbed, we intend to advance children speech recognition and related techniques. After the challenge workshop, we plan to summarize the challenge results and introduce the best-performing approaches in another follow-up paper.

## 7. REFERENCES

- [1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] Martin Russell and Shona D’Arcy, “Challenges for computer recognition of children’s speech,” in *Workshop on SLT in Education*, 2007.
- [3] Sungbok Lee, Alexandros Potamianos, and Shrikanth S Narayanan, “Analysis of children’s speech. pitch and formant frequency,” *Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 3194–3194, 1997.

- [4] Sungbok Lee, Alexandros Potamianos, and Shrikanth S Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.
- [5] Alexandros Potamianos, Shrikanth Narayanan, and Sungbok Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [6] Subrata Das, Don Nix, and Michael Picheny, "Improvements in children's speech recognition performance," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 433–436.
- [7] Qun Li and Martin J. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *International Conference on Spoken Language Processing*, 2002, pp. 2337–2340.
- [8] Sharmistha S Gray, Daniel Willett, Jianhua Lu, Joel Pinto, Paul Maergner, and Nathan Bodenstab, "Child automatic speech recognition for US english: child interaction with living-room-electronic-devices," in *Workshop on Child Computer Interaction*, 2014, pp. 21–26.
- [9] Diego Giuliani and Bagher BabaAli, "Large vocabulary children's speech recognition with DNN-HMM and SGMM acoustic modeling," in *INTERSPEECH*, 2015, pp. 1635–1639.
- [10] Piero Cosi, "A KALDI-DNN-based ASR system for Italian," in *International Joint Conference on Neural Networks*, 2015, pp. 1–5.
- [11] Romain Serizel and Diego Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," in *IEEE Spoken Language Technology Workshop*, 2014, pp. 135–140.
- [12] Hank Liao, Golan Pundak, Olivier Siohan, Melissa Carroll, Noah Cocco, Qi-Ming Jiang, Tara N Sainath, Andrew Senior, Françoise Beaufays, and Michiel Bacchiani, "Large vocabulary automatic speech recognition for children," 2015.
- [13] Mengjie Qian, Ian McLoughlin, Wu Quo, and Lirong Dai, "Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM," in *International Symposium on Chinese Spoken Language Processing*, 2016, pp. 1–5.
- [14] Joachim Fainberg, Peter Bell, Mike Lincoln, and Steve Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *INTERSPEECH*, 2016, pp. 1598–1602.
- [15] Yao Qian, Xinhao Wang, Keelan Evanini, and David Suendermann-Oeft, "Improving DNN-based automatic recognition of non-native children speech with adult speech," in *Workshop on Child Computer Interaction*, 2016, pp. 40–44.
- [16] Rong Tong, Lei Wang, and Bin Ma, "Transfer learning for children's speech recognition," in *International Conference on Asian Language Processing*, 2017, pp. 36–39.
- [17] Marco Matassoni, Roberto Gretter, Daniele Falavigna, and Diego Giuliani, "Non-native children speech recognition through transfer learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 6229–6233.
- [18] Prashanth Gurunath Shivakumar and Panayiotis Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," 2020, pp. 101077–101077.
- [19] 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 *INTERSPEECH*, 2016, pp. 2751–2755.
- [20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [21] Keyu An, Hongyu Xiang, and Zhijian Ou, "CAT: A CTC-CRF based ASR toolkit bridging the hybrid and the end-to-end approaches towards data efficiency and low latency," in *INTERSPEECH*, 2020.
- [22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, and *et al*, "The KALDI speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, 2011.
- [23] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [24] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5884–5888.

- [25] Jie Li, Xiaorui Wang, Yan Li, et al., “The speechtransformer for large-scale mandarin chinese speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 7095–7099.
- [26] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney, “A comparison of transformer and LSTM encoder decoder models for ASR,” in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 8–15.
- [27] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, and *et al*, “A comparative study on transformer vs RNN in speech applications,” in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 449–456.
- [28] Ke Hu, Antoine Bruguier, Tara N Sainath, Rohit Prabhavalkar, and Golan Pundak, “Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models,” *arXiv preprint arXiv:1906.09292*, 2019.
- [29] Tomohiro Nakatani, “Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration,” in *INTERSPEECH*, 2019, pp. 1408–1412.
- [30] Hongyu Xiang and Zhijian Ou, “CRF-based single-stage acoustic modeling with CTC topology,” in *Proc. ICASSP*, 2019, pp. 5676–5680.
- [31] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al., “ESPNET: End-to-end speech processing toolkit,” *arXiv preprint arXiv:1804.00015*, 2018.