INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

The TAL system for the INTERSPEECH2021 Shared Task on Automatic Speech Recognition for Non-Native Childrens Speech

Gaopeng Xu, Song Yang, Lu Ma, Chengfei Li, Zhongqin Wu

TAL Education Group, Beijing, China
{xugaopeng, yangsong1, malu6, lichengfei, wuzhongqin}@tal.com

Abstract

This paper describes TAL's system for the INTERSPEECH 2021 shared task on Automatic Speech Recognition (ASR) for non-native children's speech. In this work, we attempt to apply the self-supervised approach to non-native German children's ASR. First, we conduct some baseline experiments to indicate that self-supervised learning can capture more acoustic information on non-native children's speech. Then, we apply the 11-fold data augmentation and combine it with data clean-up to supplement to the limited training data. Moreover, an in-domain semi-supervised VAD model is utilized to segment untranscribed audio. These strategies can significantly improve the system performance. Furthermore, we use two types of language models to further improve performance, i.e., a 4-gram LM with CTC beam-search and a Transformer LM for 2-pass rescoring. Our ASR system reduces the Word Error Rate (WER) by about 48% relatively in comparison with the baseline, achieving 1st in the evaluation period with the WER of 23.5%.

Index Terms: children's speech recognition, self-supervised, data augmentation

1. Introduction

Despite the most recent research [1–5] indicates that the Automatic Speech Recognition (ASR) system of adult speech can reach the level close to human beings, it is still facing many challenges for non-native children's ASR. A few ASR frameworks were proposed for non-native children in [6–8] . However, there are still some difficulties and deficiencies in the application of the ASR system to these specific groups. The ASR performance will be significantly degraded by some extraordinary phenomena in non-native children's speech, including mispronounced words, grammatical errors, unsmooth words and code-switch words. Non-native children's ASR can be challenging due to language differences (e.g., acoustics, prosody, vocabulary and syntax), physiological differences (e.g., shorter vocal tract length) and cognitive differences (e.g., different levels of language cognitive ability). In addition, due to the lack of a public dataset of non-native children's speech, it is more challenging to develop an ASR system for this field. Although there are many difficulties in practical application, most of the speech transcribed by the ASR system may come from non-native speakers (such as news broadcast, movies, Internet video) and children (e.g., educational applications, voice game devices, etc.). Therefore, it is imperative to improve the ASR system for non-native speakers to process the speech of these groups accurately.

Another critical application field is the automatic assessment of children's second language proficiency. In this field, children's second language speakers have low language proficiency and lack training data. In order to deal with the lack of children's speech data, the existing research mainly focuses on children's data segmentation and the use of adult speech corpus. Speed perturbation is the most common and effective data augmentation method. Vocal track length normalization (VTLN) [9] is a standard method to alleviate the acoustic mismatch between children and adults. Recently, some methods based on deep learning have been used in children's speech recognition. However, this kind of method usually needs to use a large number of labelled children's speech training data.

In order to promote the research of ASR technology for non-native children, interspeech2021 proposes a task of the INTERSPEECH 2021 Shared Task on Automatic Speech Recognition for Non-native Children's Speech. The shared task organizers distributed a series of dataset, including English and German for non-native children speech. Datasets are generated in the context of English and German proficiency examinations.

In this paper, we detail our proposed system of a non-native German children's ASR system. We used 64h untranscribed data to train a wav2vec2.0 [10] pre-training model and about 4.6h transcribed data for fine-tuning. We also augmented the training data by manipulating the original untranscribed and transcribed speech data to simulate different speed, pitch, volume and environmental noise effects. An in-domain semi-supervised Voice Activity Detection (VAD) model is designed and utilized. Two language models (LMs) were employed in our German ASR system, i.e., a 4-gram LM constructed using the SRILM toolkit [11] and a transformer LM [12–14] used for rescoring.

The rest of this paper is organized as follows. Section 2 contains a brief description of the challenge and the baseline system. Section 3 introduces our proposed system in the non-native German ASR track. Experimental results are reported in Section 4, while conclusions is given in Section 5.

2. Baseline

2.1. Dataset

The training dataset provided in the ETLT-2021 Challenge is two parts for the German track, i.e., the TLT2017train and the deTLT1618UNTRStrain. The TLT2017train contains about 4.69 hours, manually transcribed, coming from 2017 German TLT evaluation campaigns recordings. The deTLT1618UNTRStrain contains 63.91 hours of untranscribed data, coming from 2016 and 2018 recordings. In addition, the deTLT2017dev, containing about 1.11 hours, manually transcribed, coming from 2017 recordings, was provided as dev set. The deTLT2017eval was provided as an eval set. Data statistics, such as audio duration, the number of utterances, are presented in Table 1. The ages range of speakers in the dataset is from 9 to 16 years, belonging to different school grade levels. The speakers are divided into three age-based groups, i.e., A1 (9-

http://dx.doi.org/10.21437/Interspeech.2021-1104

Table 1: *The table reports statistics for the German datasets used in our paper: two training sets: tdeTLT2017train and deTLT1618UNTRStrain, the development set (deTLT2017dev) and the evaluation set (deTLT2017eval)*

Corpus	Data Type	Duration	Utt.
deTLT1618UNTRS -train	Untranscribed	63.91h	10047
deTLT2017train	Transcribed	4.69h	1145
deTLT2017dev	Transcribed	1.11h	339
deTLT2017eval	Transcribed	1.12h	329

10), A2 (12-13) and B1 (14-16). Each age-group is asked to answer questions according to their language skills.

2.2. Baseline

A Kaldi [15] ASR system was provided as a German ASR baseline. The baseline ASR system use TLT2017train and deTLT1618UNTRStrain to train the acoustic model (AM). Firstly, GMM-HMM [16] acoustic model was trained using the TLT2017train, TDNN-F [17] models based on the labels aligned by HMM-GMM was trained using TLT2017train with 3-way speed perturbation (speed factors 0.9, 1.0 and 1.1). The input features of the TDNN-F model were 40-dim MFCC and 100-dim i-vectors. The model consists of 13 TDNN-F layers, and the context of the model is 29 frames before and after the underlying frame, i.e., [-29, +29]-frames. An n-gram language model (LM) was trained on the 2016 and 2018 campaigns' written answers. Then, this preliminary model was used to create alignments for the deTLT1618UNTRStrain. The final TDNN-F model was trained with a weighted combination of supervised and unsupervised data, and the procedure follows the recipe proposed in the Kaldi recipe.

We also built baseline models including chain, speech-transformer and wav2vec2.0 (unsupervised pre-training). In all BASE models, we used TLT2017train with 3-way speed perturbation as training data. The deTLT1618UNTRStrain dataset was used in the chain model for Semi-Supervised training and wav2vec pre-training.

For the chain model, we used the multi-stream self-attention (MSA) [18] model structure instead of the TDNN-F. This architecture allows input speech frames to be efficiently processed and effectively self-attention, getting better performance. The configuration specified was 4 layers of MSA blocks, 3 streams of self-attention in each block, 8 layers of TDNN-F in each stream. The dimension of Self-attention was set to be 40 (20 heads). Finally, we used the Minimum Bayes-Risk (MBR) decoding algorithm as proposed in [19]. This BASE models also made use of data clean-up and resegment as is described in section 3.2.

For the Transformer model, a 12-layer encoder and 6-layer decoder was used, and the dimension of attention and the feed-forward layer was set to be 512 (8 heads) and 2048, respectively. As for the pre-training model, an unsupervised training Wav2Vec2.0 encoder was used. We used the deTLT1618UNTRStrain dataset for unsupervised training and the TLT2017train dataset for supervised training.

Table 2 shows the WERs for the ASR systems trained using different BASE models. In the baseline experiments, we found that the pretraining-BASE models achieved the best performance. The chain model got the second-best result. Since the transformer model just used 4.69h supervised data, and the amount of data was tiny, the result was deplorable.

Table 2: % *WER of baseline system on Dev set, trained on TLT2017train with 3-way speed perturbation*

Acoustic model	Data Size	WER (%)
Chain	3x	45.3
Transformer	3x	68.8
Pre-training	3x	43.5

3. System Description

3.1. Data Augmentation

To increase the amount of training data and enhance the robustness of the model against different speaking styles, reverberation, and background noises, we applied 7 types of data augmentation methods, namely speed perturbation (speed factors 0.9,1.0 and 1.1), volume perturbation (volume factors 0.9 and 1.1), reverberation simulation, babble noise augmentation, music noise augmentation, non-speech noise augmentation and pitch augmentation (pitch scale 0.85 and 0.9) [20, 21]. These types of data augmentation, in total, resulted in an 11-fold increase for the training data.

3.2. Resegment and Data clean-up

We observed that the original manual transcriptions were provided with symbols representing speech and non-speech noise events such as laughter and noise, but the transcriptions were not wholly accurate. For example, some audio duration is about 30 seconds, but the annotation result is only one word, and a little audio contains only noise or silence that is not friendly for training acoustic models. In addition, some of the original audio is longer than 100s, which affects the efficiency of using GPU to train the model. We train a chain-based model to re-segment and clean-up training data, selecting the good audio that matches the transcripts. We decode the original data with a pretraining in-domain chain-base acoustic model and a biased language model built from the reference transcript and then obtained the segmentation from a CTM-like file. We get about 2.6h of re-segment data, and the maximum length of the audio clip is 30s. Because the amount of training data is relatively small, we finally combine the data less than 50s in the original data with the data after clean-up to get about 6.7 hours of supervised training data.

3.3. Pre-training

For the pre-training model, an unsupervised training Wav2Vec encoder was used, it consists of a feature encoder, context network and quantization module. We performed experiments with BASE model and LARGE model, which used the same encoder but differ in the Transformer block. The BASE model feature encoder contains 7 convolutions blocks, each block with strides (5,2,2,2,2,2,2) and kernel widths (10,3,3,3,3,2,2). The number of output channels is 512. Context network contains 12 transformer blocks, each block with 512-dim model, 8 attention heads, and 2048-dim feed-forward inner-layer. The LARGE model contains 20 layers of transformer block with 16 heads, and the hidden dimension is 1024, and the final dimension (FFN) is 4096. Since there are some long audios in the datasets

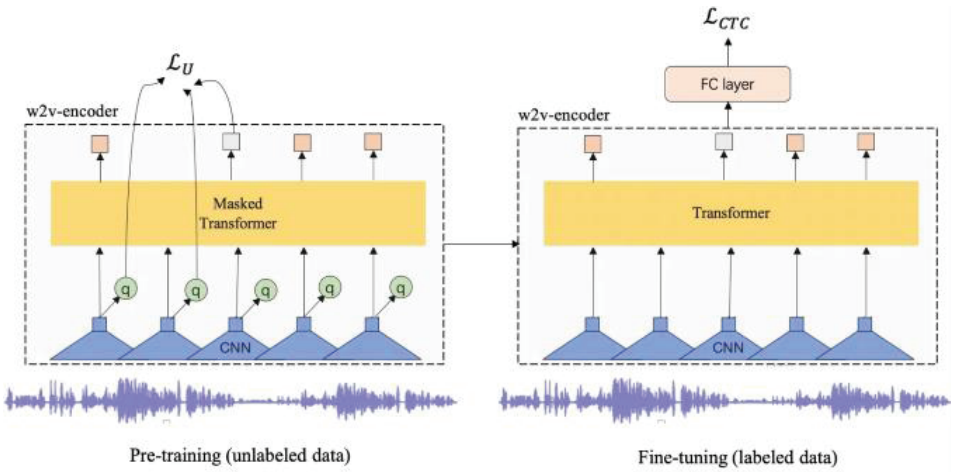

Figure 1: *The process of the pre-training and fine-tuning.*

provided by the organizer, we cut the long audio into short segments with no more than 30s and trained the baseline model on 8 number of V100 GPUs. The model was optimized with Adam. During the first 10% updates, the learning rate warms up to a peak of 3×10^{-5}, and then it decays linearly. The process of the pre-training and fine-tuning is illustrated in Fig. 1.

3.4. Semi-supervised VAD model

Considering that the long audio may not be cut into 30-second segments appropriately, we used the Semi-supervised Voice Activity Detection (VAD) model to segment the original audio. We decoded untranscribed audio with an existing in-domain chain model mentioned in section 2.2. VAD targets were obtained by decoding. The phones in the lattices were mapped to 0, 1, and 2, representing silence, speech and garbage classes,respectively. These targets were used to train a 4-layer TDNN network for VAD. We used this VAD model in the ASR system as will be described in section 4, obtaining considerable gains.

3.5. Fine-tuning

For fine-tuning, We use a total of 35 tokens, including 30 German characters, a word boundary token and four non-linguistic Sharing symbols (i.e., @noise, @unk, @sil, @hes), as is shown in Table 3. Models are optimized by minimizing CTC loss, and we also apply masking to time-steps and channels during fine-tuning, which reduces overfitting and improve the recognition robustness like Specaugment [22]. We fine-tuned on the transcribed data TLT2017train with pre-training model and added a randomly initialized output layer on top of the Transformer to predict characters. The model was optimized with Adam, and the learning rate was warmed-up for the first 1k of steps. We used a batch size of 1.28m samples per GPU, and we fine-tuned on 8 number of V100 GPUs. For the first 1000 updates, only the final output classifier was trained, after which the Transformer block was also trained. The feature encoder was frozen during fine-tuning training. The learning rate was set to 3×10^{-5}, and the model with lowest WER on dev set was chosen as the checkpoints, which was decoded with the 4-gram LM with 100 beam-width, where the LM weight is 2 and the word insertion penalty was set to be -1.

Table 3: *Sharing the Non-linguistic symbol*

Non-linguistic symbols	**Sharing symbols**
@a @ae @ah @am @am @e @eem @ef @ehh @em @eo @er @f @ge @h @hes @hm @i @m @mm @mh @o @oh @s @uh @um	@hes
@bkg @boh @breath @cough @laugh @ns @noise	@noise
<unk-de> <unk-en> <unk-it> unk @voice @voices	@unk
@sil	@sil

3.6. Language Model

Two n-grams LM were used in our system. A 4-gram LM trained by the TLT2017train transcript and a 4-gram LM trained by transcribing the detlt1618untrrain with the chain-based model. During decoding, we used a certain confidence threshold to remove some bad results. We linearly interpolated these models trained on different text, which was improved on the single models (LM weight 2, word insertion penalty -1). We also used Transformer LM for 2-pass rescoring. The Transformer LM contains 12 blocks with 16 heads, and the dimension of attention and the feed-forward layer was set to 512 and 2048, respectively.

4. Results

Table 4 shows the results of different models starting from the baseline model to the best performance model. As baseline pre-training, we used the deTLT1618UNTRStrain dataset containing 64 hours of unlabeled audio, and we followed section 2.2 to train the pre-training model. After pre-training, we fine-tuned the learned representations on the TLT2017train containing 4.69 hours of labelled data and added an output layer on top of the Transformer to predict characters and the model were optimized by minimizing CTC loss. The mode was optimized with Adam and a learning rate of 3×10^{-5}. The WER on the development set of the baseline model was 43.5%.

Model 1 shows the improvement gained by 11-fold data augmentation in transcribed which are speed perturbation

Table 4: % *WER of the dev sets for different model with data augmentation (speed perturbation, volume perturbation, reverberation simulation, babble augmentation, music augmentation, noise augmentation and pitch augmentation), data clean-up, VAD segment(vad_seg) and language model*

ID	Model	Pretraing Data	Fine-tuning Data	Pretraing	Fine-tuning	LM	WER (%)
Baseline	BASE model	deTLT1618UNTRStrain	TLT2017train 3x	64h	13.8h	no	43.5
Model 1	BASE model	deTLT1618UNTRStrain	TLT2017train 11x	64h	51.6h	no	38.9
Model 2	BASE model	deTLT1618UNTRStrain 11x	TLT2017train 11x	704h	51.6h	no	34.1
Model 3	BASE model	deTLT1618UNTRStrain 11x	TLT2017train +clean-up 11x	704h	73.59h	no	30.5
Model 4	BASE model	deTLT1618UNTRStrain +vad_seg 11x	TLT2017train +clean-up 11x	1166h	73.59h	no	29.6
Model 5	LARGE model	deTLT1618UNTRStrain +vad_seg 11x	TLT2017train +clean-up 11x	1166h	73.59h	no	28.2
Model 6	LARGE model	deTLT1618UNTRStrain +vad_seg11x	TLT2017train +clean-up 11x	1166h	73.59h	4-gram	27.6
Rescoring of the decoded 100-best of model5 with transformer lm							26.8

(speed factors 0.9 and 1.1), volume perturbation (volume factors 0.9 and 1.1), reverberation simulation, babble noise augmentation, music noise augmentation, non-speech noise augmentation and pitch augmentation (pitch scale 0.85 and 0.9). A 10.5% decrease in the WER was achieved by increasing the transcribed training data augmentation.

Model 2 shows the improvement gained by increasing the transcribed training data to around 51.6 hours and untranscribed training data to around 704 hours using 11-fold data augmentation. A 11.4% decrease in the WER was achieved by increasing both transcribed and untranscribed training data.

We checked the data provided by the challenge organizers and found that some transcribed were not completely accurate. We follow section 2.3 to resegment, and clean up transcribed data. Then, we combined the data shorter than 50s in the original data with the data after clean-up to get about 6.7 hours of transcribed training data. In Model 3, we also used 11-fold data augmentation for 6.7 hours of combined data, and the WER dropped from 34.1% to 30.5%.

In Model 4, we used the VAD model mentioned in section 2.3 to segment the original untranscribed audio and the duration is from 64h to 42h. To ensure that there are enough available pre-training data, we combined the original segments and the segments through the VAD model to obtain 106 hours of data and used 11-fold data augmentation to obtain around 1166 hours. The Model 5 is like Model 4, but uses LARGE model instead of BASE model for pre-training and the WER dropped from 29.6% to 28.2%.

In Model 6, a 4-Gram language model was joined to decode based on Model 5, and the model achieved a WER of 27.6%. The output 100-best of Model 5 was further rescored using Transformer LM and gave 0.8% reduction in WER.

Table 5 shows the result submitted for the German track, giving a WER of 23.5% on the eval set, which was generated by Model 6 with transformer-LM rescoring.

5. Conclusions

In this paper, we presented TAL's ASR system for the task of recognizing non-native German children's speech. We focused on applying a self-supervised-based approach. Compared to supervised training, the self-supervised-based model achieved better results in this limited training data task. We employed the 11-fold data augmentation, data clean-up and Semi-supervised VAD to improve the ASR performance. A 4-gram language model with CTC beam search estimated from children's transcribed speech was used for the decoding step, and the final step was performed by rescoring 100-best using Transformer LM. Our best-submitted systems rank in the first place for the German closed and open track.

Table 5: % *WER of best-submitted system on Eval set*

Acoustic model	Language model	Test	WER (%)
Model 5	4-gram +transformer-LM	Eval	23.50

6. Acknowledgements

This work was supported by National Key R&D Program of China, under Grant No. 2020AAA0104500.

7. References

[1] L. Dong, S. Xu and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884-5888.

[2] Y. zhao, J. Li, X. Wang and Y. Li, "The Speechtransformer for Large-scale Mandarin Chinese Speech Recognition," *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7095-7099.

[3] K. Anjuli, D. Arindrima, N. Tara, "Large-scale multilingual speech recognition with a streaming end-to-end model," in *Proc. ISCA Interspeech*, 2019, pp. 2130–2134.

[4] A. Gulati, J. Qin, C. C. Chiu, et al. "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint*, arXiv:2005.08100, 2020.

[5] K. J. Han, J. Pan, V. K. N. Tadala, et al. "Multistream CNN for robust acoustic modeling," *arXiv preprint*, arXiv:2005.10470, 2020.

[6] K. Evanini and X. Wang, "Automated speech scoring for nonnative middle school students with multiple task types," in *Proc. of INTERSPEECH*, 2013, pp. 2435-2439.

[7] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional LSTM-RNN for Improving Automated Assessment of Non-native Children's Speech," in *Proc. of INTERSPEECH, 2017*, pp. 1417–1421.

[8] M. Matassoni, R. Gretter, D. Falavigna and D. Giuliani, "Non-Native Children Speech Recognition Through Transfer Learning," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6229-6233.

[9] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 135-140.

[10] A. Baevski, Y Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Neural Information Processing Systems (NeurIPS)*, 2020.

[11] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Interspeech 2002*, 2002.

[12] H. Huang, F. Peng. "An Empirical Study of Efficient ASR Rescoring with Transformers," *arXiv preprint*, arXiv:1910.11450, 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al. "Attention is all you need," in *Advances in neural information processing systems, 2017*, pp. 5998–6008.

[14] K. Irie, A. Zeyer, R Schlüter, et al. "Language modeling with deep transformers," in *INTERSPEECH, 2019*.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek,N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[16] P. Bansal, A. Kant, S. Kumar, et al. "Improved hybrid model of HMM/GMM for speech recognition," in *ITHEA 2008*, pp. 69-74.

[17] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of INTERSPEECH, 2018*, pp. 3743–3747, 2018.

[18] K. J. Han, R. Prieto, and T. Ma, "State-of-the-art speech recog- nition using multi-stream self-attention with dilated 1D convolution," in *ASRU, 2019*, pp. 54–61.

[19] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," in *Computer Speech & Language*, vol. 25, 2011.

[20] W. Ahmad, S. Shahnawazuddin, H. Kathania, G. Pradhan, and A. Samaddar, "Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion," in *Interspeech 2017*, pp. 2391–2395.

[21] H. K. Kathania, W. Ahmad, S. Shahnawazuddin, and A. B.Samaddar, "Explicit pitch mapping for improved children's speech recognition," in *Circuits, Systems, and Signal Processing*, vol. 32, pp. 2021–2044, 2018.

[22] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le. "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. of Interspeech 2019*.