



# Leveraging Phone Mask Training for Phonetic-Reduction-Robust E2E Uyghur Speech Recognition

Guodong Ma<sup>1</sup>, Pengfei Hu<sup>2</sup>, Jian Kang<sup>2</sup>, Shen Huang<sup>2\*</sup>, Hao Huang<sup>1\*</sup>

<sup>1</sup>School of Information Science and Engineering, Xinjiang University, Urumqi, China

<sup>2</sup>Tencent Minority-Mandarin Translation, Beijing, China

{mag431217, hwanghao}@gmail.com, studenthu@163.com, kangjianqh@sina.com, qqhuangshen@hotmail.com

## Abstract

In Uyghur speech, consonant and vowel reduction are often encountered, especially in spontaneous speech with high speech rate, which will cause a degradation of speech recognition performance. To solve this problem, we propose an effective phone mask training method for Conformer-based Uyghur end-to-end (E2E) speech recognition. The idea is to randomly mask off a certain percentage features of phones during model training, which simulates the above verbal phenomena and facilitates E2E model to learn more contextual information. According to experiments, the above issues can be greatly alleviated. In addition, deep investigations are carried out into different units in masking, which shows the effectiveness of our proposed masking unit. We also further study the masking method and optimize filling strategy of phone mask. Finally, compared with Conformer-based E2E baseline without mask training, our model demonstrates about 5.51% relative Word Error Rate (WER) reduction on reading speech and 12.92% on spontaneous speech, respectively. The above approach has also been verified on test-set of open-source data THUYG-20, which shows 20% relative improvements.

**Index Terms:** Uyghur, speech recognition, end-to-end ASR, Conformer, Phone Mask

## 1. Introduction

In recent years, with the rapid development of neural network, the performance of ASR system has been greatly improved. Current traditional ASR systems consists of several components including acoustic, lexicon, and language models. E2E ASR systems train these models jointly and the systems map acoustic feature sequences to token sequences directly. E2E ASR framework can generally be divided into the following categories: connectionist temporal classification (CTC) [1, 2] and attention-based sequence-to-sequence encoder-decoder [3, 4, 5]. In this context, many works are based on E2E frameworks to explore different ASR scenarios and improve the performance accordingly. For example, code-switch [6, 7, 8, 9] and low-resource task [10, 11, 12] etc.

The Uyghur language, which belongs to Altaic language family, is mainly spoken by a large population in Xinjiang, China. Research on Uyghur ASR has been carried out for many years. The study on ASR of the Altaic language family generally focuses on issues, such as agglutinative characteristics and low resources [13, 14, 15, 16, 17, 18, 19]. Altaic languages, such as Uyghur, is spoken much faster than other languages, such as Mandarin, English, etc. Therefore, in the daily life of Uyghur, there will be common language phenomena such

as consonant and vowel reduction, which is referred as phonetic reduction (PR). The phones which have undergone PR become shorter. Accordingly, it brings great difficulties to the construction of robust Uyghur ASR. In the conventional ASR, PR phenomena can be straightforwardly modeled by using a multi-pronunciation dictionary, in which the possible PR pronunciation variants are included. But it is difficult to get all possible pronunciation variants normally, and need require the knowledge of language experts. In E2E ASR systems, the learning of pronunciation dictionary information is integrated in the model, so how to modify the model framework to learn dictionary information is a key question to explore.

We try to solve PR from a mask training perspective under the E2E ASR framework. Recently, a lot of excellent works about masking have been proposed. BERT [20] masks the modeling units so that the model can consider more information of the context. SpecAugment [21] masks the spectrum of time or frequency domain in random. It increases the diversity of training data to prevent over-fitting. In addition, semantic mask [22] randomly masks spectrum of word of 15% during training E2E model. It aims to improve the strength of LM. However, the lack acoustic meaning of masking, the random and large masking size, make above algorithms cannot be connected with the problems we need to solve directly. Inspired by the above masking methods [20, 21, 22] and the peculiarities in Uyghur, we propose an effective masking method, phone mask (PM). The main idea is to randomly mask off a certain percentage features of phones during model training. Having such case in Uyghur of which a phone stands for a grapheme (except individual phone), so phone mask generally equal to grapheme mask. It simulates the above verbal phenomena and facilitates model to learn more contextual information. In addition, we further to discuss the masking ways and filling strategies of phone mask. Seeing section 3 for details. Compared with E2E model baseline without masking training, our model indicates an WER relative reduction of 5.51% on reading speech and 12.92% on spontaneous.

Our main contributions are as follows: To our best knowledge, we are the first to use a large amount of Uyghur speech data to explore ASR task under the E2E framework. And also perhaps we are the first to address the issue of PR and provide a solution to this problem under the Conformer based E2E ASR framework. Moreover, through experiments and investigations into a variety of masking configurations are presented.

## 2. Related Work

Recent research on ASR of Altaic languages such as Uyghur are mainly focusing on the solution to the OOVs problem caused by the stickiness and under-resourced problem caused by data scarcity [13, 14, 15, 16, 17, 18, 19]. To our best knowledge,

\* corresponding authors

there were no relevant papers that had researched the solutions to the problems of PR for ASR. Recently, emerging masking methods such as BERT [20], SpecAugment [21] and semantic mask [22] has shown to be very successful. SpecAugment [21] focuses more on the diversity of data onto model training, so as to prevent to over-fitting. The other two masking strategies [20, 22] concern more about utilizing the surrounding information as much as possible to learn abundant LM knowledge and improve the robustness of model. The PR occurrence depends greatly on the contextual pronunciations and might be learned through large amount speech data. Therefore, application of the mask training to solving the PR problem in Altaic language, such as Uyghur, can be reasonable and feasible.

### 3. Proposed Method

#### 3.1. Phonetic reduction in Uyghur speech

Phonetic reduction are very common in Altaic languages<sup>1 2</sup>, especially in high speed spontaneous speech. To give a clearly illustration, we built a conventional hybrid ASR model using Kaldi [23] and compute length of each phone by doing phone-level forced alignment. Firstly we compare the average length of phonemes in the languages of Mandarin, English and Uyghur. As shown in Table 1, the average duration of Uyghur phones are obviously shorter than that of Mandarin and English, while Mandarin has the longest phone duration.

According to analysis of the forced alignment results, Uyghur has higher ratios of phones which duration is 3 frames. Since the HMM models used in force alignment has 3 states, every phone has 3 frames duration at least during force alignment. We check these short duration phones' speech and find that many of them has phonetic reduction phenomenon to some extent. It is also found that the high speaking speed aggravate phonetic reduction in Uyghur.

Table 1: The average duration of phonemes

Language	Duration (Seconds)
Mandarin	0.14
English	0.10
Uyghur	0.08

#### 3.2. Masking Training

The masking training of E2E speech recognition can be traced back to SpecAugment [21], which improved the system robustness by masking spectrum randomly in time or frequency domain. Instead of randomly masking, the semantic mask method masks the acoustic features corresponding to a particular word, which improves the power of the implicit language model in the decoder. Aiming at reducing the impact of phonetic reduction in Uyghur speech, we propose the phone mask method, which masks the acoustic features of particular phone. To achieve this, we need to obtain alignment information on the training data. As shown in Figure 1, we train an HMM-DNN (nnet3) model using Kaldi and get the alignment information of each phone and frame using forced-alignment. During model training, we randomly select a percentage of the phones and mask the corresponding speech segments in each iteration as shown in Fig-

<sup>1</sup>[https://en.wikipedia.org/wiki/Altaic\\_languages](https://en.wikipedia.org/wiki/Altaic_languages)

<sup>2</sup>[https://en.wikipedia.org/wiki/Uyghur\\_language](https://en.wikipedia.org/wiki/Uyghur_language)

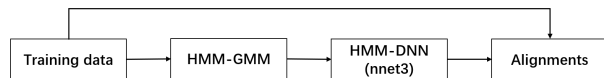


Figure 1: Get alignments using kaldı

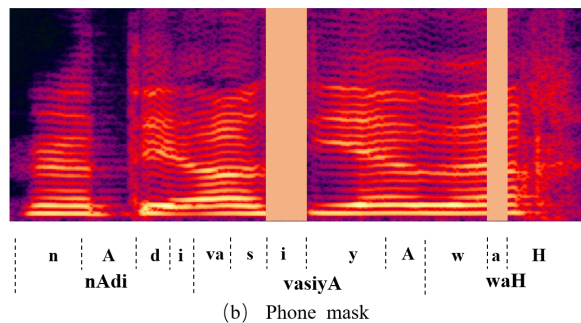
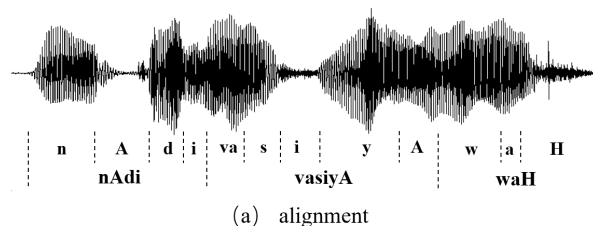


Figure 2: Example of alignment and phone mask

ure 2. In experiment part later, we make a deep study about masking ratio and the filling strategy.

The proposed phone mask training is to some extent similar to semantic mask. They both mask spectrum during model training. But the biggest difference between our method and semantic mask [22] is that our method not only to encourage E2E model to learn more contextual information, but also to make it to learn more dictionary knowledge. The intuitions behind these methods are also different. Our proposed phone mask aims to phonetic reduction characteristic of the Uyghur speech. Meanwhile, we think that the word-level semantic masking has too large masking span and will put great pressure for the model learning. We will experiment with different masking granularity later to prove this conjecture.

## 4. Model

### 4.1. Conformer Encoder

Conformer [5] is a SOTA ASR encoder architecture, which inserts a convolution layer into transformer block to increase the local information modeling capability of the traditional Transformer model. The architecture of Conformer encoder consists of a self-attention, convolution, feed-forward, and layernorm modules. The second feed-forward module is followed by a final layernorm layer. For example  $x$  is the input of the Conformer block  $i$ , then the  $i$ -th block output is calculated as

$$x' = x + \frac{1}{2}FFN(x) \quad (1)$$

$$x'' = x' + MHSA(x') \quad (2)$$

$$x''' = x'' + Conv(x'') \quad (3)$$

$$y_i = Layernorm(x''' + \frac{1}{2}FFN(x''')) \quad (4)$$

where  $\text{FFN}(\cdot)$ ,  $\text{MHSA}(\cdot)$  and  $\text{Conv}(\cdot)$  denote the Feed-Forward Network Module, Multi-Head Self-Attention Module and Convolution Module.

## 4.2. ASR Training and Decoding

Following [24, 25], we use Conformer (CE) and CTC objectives for training and decoding. In the training step, all models are trained to minimize both Conformer (CE) and CTC losses jointly. The objective function is described as follows:

$$\text{Loss} = -\alpha \text{Loss}_{ce} - (1 - \alpha) \text{Loss}_{ctc} \quad (5)$$

where  $\alpha$ , a tunable parameter, denotes an weight that balances the Conformer loss (CE loss) and CTC loss. In this paper, we set  $\alpha = 0.7$ . In the decoding stage, we combine the probabilities of Conformer with CTC as follows:

$$\hat{Y} = \arg \max_y (\lambda \log P_{ce}(y|X) + (1 - \lambda) \log P_{ctc}(y|X)) \quad (6)$$

where  $X$  represents the inputs and  $\lambda$  is a tunable parameter. Following [22], we set  $\lambda$  to 0.5.

## 5. Experiments and Results

### 5.1. Experimental setup

For the training data, we collect 1200 hours Uyghur speech corpus from several data companies. As for the evaluation set, we use speech datasets collected from two scenarios, one is reading speech (Read-Test) and the other is spontaneous speech (Oral-Test). Among them, the speech rate of Oral-Test set is much faster than the Read-Test set. The details about the datasets are shown in Table 2. All the experiments use 40 Mel Frequency Cepstral Coefficients (MFCCs) over 25 ms frames with 10 ms stride to each of which cepstral mean and variance normalization (CMVN) is applied. We both experiment using the state-of-the-art hybrid and E2E ASR systems. We add 100 i-Vector only to the hybrid ASR system for speaker adaptation.

Table 2: Descriptions of the datasets

Data Type	Duration (Hours)	Domain	Avg Phone Dur. (Sec.)
Train	1200	Reading/Clean	0.09
Read-Test	3	Reading/Clean	0.10
Oral-Test	5	Oral/Noisy	0.06

For the hybrid ASR system, we adopt the chain model [26] using Kaldi toolkit [23]. The neural network has 15 hidden layer TDNN and a rank reduction layer. The number of units in TDNN consists of 1563 and 160 bottleneck unit. The learning rate for training with 5 epochs is chosen initially from 0.00025 ending at 0.000025. We have developed 4-gram LM using the transcription of training data trained on SRILM toolkit [27].

As for the E2E system, we adopt the Conformer [5] and Transformer architecture of [22], named STM-Transformer. For comparability, the configurations of our Conformer are the same as STM-Transformer (Encoder = 12, Decoder = 6, H = 8,  $d^{att} = 512$ ). Other configurations of Conformer follow [25]. The output tokens are 5000 BPE tokens produced by unigram model using sentencepiece [28]. All the E2E models are trained by using the speech recognition toolkit ESPnet [29] on 4 P40 GPUs, which are trained 40 epochs (decreased epochs due to speed perturbation [30]) spending about 5 days.

Table 3: Comparisons between semantic mask (STM), word piece mask (WPM), phone mask (PM), and PM variants using Conformer[5]; Mask methods\_(number) mean the mask ratio is number, if have no number, 15 % by default. (WER)

SYSTEM	Read-Test	Oral-Test
Chain	26.0	47.9
Grapheme-based Conformer	26.8	45.0
Word-piece-based Conformer	25.4	44.1
+ SpecAugment	25.2	41.6
+ STM	25.1	42.4
+ WPM	25.0	40.1
+ PM	24.8	38.9
+ WPM_20	24.4	40.0
+ PM_20	24.2	38.9
+ PM_20 + _FW	<b>24.0</b>	<b>38.4</b>

### 5.2. Results

#### 5.2.1. Results on Conformer

To our best knowledge, no related work before explored the E2E Uyghur speech recognition using a large amount of speech data, so we firstly conduct the experiments on grapheme-based and word-piece-based Conformer and compare the results with the chain model. It can be seen that the Conformer based system show better performance than the chain model on the Oral-Test, but the grapheme-based Conformer show slightly worse performance than the chain model on the Read-Test. The reason might be that the transcripts used in LM of chain model and Read-Test set belong to similar domains, which enhances the performance of Read-Test with Chain model, while grapheme-based Conformer’s decoder is hard to model the relationship across words. Similar to general E2E ASR, word-piece-based Conformer outperforms chain model and grapheme-based Conformer on both of the two test sets.

For mask training, we firstly experiment on different masking methods including SpecAugment, semantic mask (STM), and phone mask (PM). By comparing the results of SpecAugment with STM, it can be observed that the former performs better on the Oral-Test set. We think it is because the word-level STM has a relative larger masking granularity, which put great overhead on model learning. However, SpecAugment, especially time masking, is capable of alleviating the consonant and vowel reduction in Uyghur to some extent. Overall, our proposed PM, which is inspired by the peculiarities of Uyghur, achieves the best performance among them, which show the effectiveness for phonetic reduction and further demonstrate the granularity play an important role in masking. Secondly, in order to more effectively illustrate the issue of the large masking granularity of STM, we replace the word-level masking (STM) with word-piece-level, which is named word-piece mask (WPM). The results by WPM show better than those from word-level masking (STM) and SpecAugment. This is consistent with our hypothesis that it hurt the WER if the masking granularity is too large.

We also experiment with different mask ratio, i.e. the percentage of phone or word-piece that are masked in the training process. The default mask ratio is set to 15% in the above experiments, following the BERT [20] and semantic mask [22]. We further adjust the ratio of masking from 15% to 30% for the experiment. We have observed the masking ratio of 20% is the optimum for phone mask and word piece mask. The results are

also shown in Table 3, which also indicates that the granularity of word-level masking affects the effect of masking to some extent. As for filling strategy (PM\_20 + \_FW), besides filling with the average value of the whole utterance features, we also try to use the average value of word where the selected masked phone is located. It enables the model to make use of more relevant information. We can see the performance, in the last row of Table 3, has been further improved.

As mentioned in Section 3, phonetic reduction can be aggravated by high speaking speed. Therefore, it is naturally to consider the utilization of speed perturbation [30] to augment the speech data of variant speed. In Table 4, we experiment on speed perturbation, PM, and PM after speed perturbation. By comparing speed perturbation only (SP) and PM only, we can see that pure PM achieves better performance than SP on both of the two test sets. Especially, the improvements on the Oral-Test set is more obvious, which further demonstrates the effectiveness of the proposed PM training. Finally, when we do PM after SP, it shows a 2.9% absolute improvement on Oral-Test and 1.0% on Read-Test compared with SP only system.

Table 4: Comparison with speed perturbation (WER)

SYSTEM	Read-Test	Oral-Test
Word-piece-based Conformer	25.4	44.1
+ PM_20 + _FW	24.0	38.4
+ Speed Perturb	24.5	39.9
+ PM_20 + _FW	<b>23.5</b>	<b>37.0</b>

### 5.2.2. Results on the THUYG-20 benchmark's test set

To further demonstrate the effectiveness of the proposed method, we present the results on a popular open-source Uyghur ASR benchmark, THUYG-20 [31]. Table 5 shows the results. The results are in consistency with the conclusions reached above. The PM\_20 + \_FW system shows the best performance with a WER of 10.0% on the THUYG-20 test set. To better illustrate how the proposed method alleviate the problem of phonetic reduction, in Figure 3, we show a case of a THUYG-20 test utterance with its ground-truth label and the corresponding recognition outputs using different model configurations. We can see the issue of phonetic reduction can be solved by the proposed method from the perspective of masking. The results show that the phone mask we proposed can better reduce substitution errors and achieves the best WER.

Table 5: WER on the THUYG-20 test set

SYSTEM	WER	SUB	DEL	INS
Base Conformer	12.5	10.4	1.4	0.7
+ SpecAugment	12.6	10.3	1.4	0.9
+ STM	12.4	10.1	1.5	0.8
+ PM_20	10.6	8.7	1.4	0.5
+ PM_20 + _FW	<b>10.0</b>	<b>8.3</b>	<b>1.2</b>	<b>0.5</b>

### 5.2.3. Results on STM-Transformer

To provide a fair comparison and the applicability of our proposed method to other network structure, we make an experimental comparison on STM, WPM and our proposed PM using the STM-Transformer network structure and configuration proposed on [22]. Table 6 presents the results (WER). Similar to

#### Ground truth

mAn bu tomuz yazni dala lageri bazisida vOtkUzdUm

#### Conformer baseline

mAn bu tomuz yazni dala lageri bazsida kOzUm

#### Conformer with STM

mAn bu tomuz yazni dala lageri bazsida kUzdUm

#### Conformer with PM\_20 + \_FW

mAn bu tomuz yazni dala lageri bazisida vOtkUzdUm

Figure 3: Case of THUYG-20 test-set (UTT\_ID: F1010\_037)

[22], the baseline Transformer represents the model with position embedding. The proposed PM training still achieves the best performance on the two test sets. Compared with the baseline and STM, the PM\_20 model show absolute improvements of 5.2% and 3.3% on the Oral-Test set, respectively, and a significant improvement on the Read-Test set as well. The WER reduction on the Oral-Test set is greater than that on the Read-Test set. Meanwhile, we can see from the Table 6 that, both Read-Test and Oral-Test, WPM is better than STM, which further confirms our guess in Section 3 that too large masking span will put great overhead for the model learning.

Table 6: Comparison of mask training using STM-Transformer

SYSTEM	Read-Test	Oral-Test
Base Transformer	28.9	44.7
STM-Transformer	27.9	44.3
STM-Transformer with STM	25.6	42.8
STM-Transformer with WPM_20	25.1	41.7
STM-Transformer with PM_20	<b>24.5</b>	<b>39.5</b>

## 6. Conclusions

In this paper, aiming at resolving the phenomenon of phonetic reduction in Altaic language, such as Uyghur, we propose an effective phone masking method for E2E speech recognition. During training, a certain proportion spectrum of phones will be masked and filled with the average of word span spectrum features. Therefore, more contextual information and dictionary knowledge are expected to be learned with the help of phone masking. We have carried out Uyghur speech recognition tests on both reading and spontaneous speech. Results show that proposed method improves the performance of Uyghur speech recognition obviously, especially on the spontaneous speech. Extensive investigations into the masking granularity, comparison with traditional masking methods and applicability to other model structure have also been carried out and confirm the effectiveness of the proposed method. According to the analysis, the phone masking training is very helpful to improve the robustness of Uyghur speech recognition system to PR problem.

## 7. Acknowledgements

The authors of this paper would like to give special thanks to Dr. Nurmamet Yolwas with Xinjiang University for guidance and help in this work. This work was supported by the National Key R&D Program of China (2017YFB1402101) and Natural Science Foundation of China (61663044, 61761041).

## 8. References

- [1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: <https://doi.org/10.1145/1143844.1143891>
- [2] A. Graves, "Sequence Transduction with Recurrent Neural Networks," *arXiv e-prints*, p. arXiv:1211.3711, Nov. 2012.
- [3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4960–4964.
- [4] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.
- [5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [6] Y. Lu, M. Huang, H. Li, J. Guo, and Y. Qian, "Bi-Encoder Transformer Network for Mandarin-English Code-Switching Speech Recognition Using Mixture of Experts," in *Proc. Interspeech 2020*, 2020, pp. 4766–4770.
- [7] Z. Qiu, Y. Li, X. Li, F. Metze, and W. M. Campbell, "Towards Context-Aware End-to-End Code-Switching Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 4776–4780.
- [8] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Investigating end-to-end speech recognition for mandarin-english code-switching," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6056–6060.
- [9] J. MetildaSagayaMaryN., V. M. Shetty, and S. Umesh, "Investigation of methods to improve the recognition performance of tamil-english code-switched data in transformer framework," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7889–7893, 2020.
- [10] Y. Sharma, B. Abraham, K. Taneja, and P. Jyothi, "Improving Low Resource Code-Switched ASR Using Augmented Code-Switched TTS," in *Proc. Interspeech 2020*, 2020, pp. 4771–4775.
- [11] C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7719–7723.
- [12] V. M. Shetty and M. Sagaya Mary N.J., "Improving the performance of transformer based low resource speech recognition for indian languages," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8279–8283.
- [13] P. Hu, S. Huang, and Z. Lv, "Investigating the Use of Mixed-Units Based Modeling for Improving Uyghur Speech Recognition," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 215–219.
- [14] A. Abulimiti and T. Schultz, "Automatic speech recognition for Uyghur through multilingual acoustic modeling," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6444–6449.
- [15] P. Smit, S. Virpioja, and M. Kurimo, "Advances in subword-based HMM-DNN speech recognition across languages," *Computer Speech & Language*, vol. 66, p. 101158, 2021.
- [16] C. Liu, Z. Zhang, P. Zhang, and Y. Yan, "Character-Aware Sub-Word Level Language Modeling for Uyghur and Turkish ASR," in *Proc. Interspeech 2019*, 2019, pp. 3495–3499.
- [17] L. Shi, F. Bao, Y. Wang, and G. Gao, "Research on transfer learning for khalkha mongolian speech recognition based on TDNN," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 85–89.
- [18] A. Bisazza and R. Gretter, "Building a turkish ASR system with minimal resources," in *Proceedings of First Workshop on Language Resources and Technologies for Turkic Languages*, 2012.
- [19] M. Ablimit, G. Neubig, M. Mimura, S. Mori, T. Kawahara, and A. Hamdulla, "Uyghur morpheme-based language models and asr," in *International Conference on Signal Processing Proceedings*, 11 2010, pp. 581 – 584.
- [20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv e-prints*, p. arXiv:1810.04805, Oct. 2018.
- [21] 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. Interspeech 2019*, 2019, pp. 2613–2617.
- [22] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, "Semantic Mask for Transformer Based End-to-End Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 971–975.
- [23] 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," *Idiap, Rue Marconi 19, Martigny, Idiap-RR-04-2012*, 1 2012.
- [24] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning," *arXiv e-prints*, p. arXiv:1609.06773, Sep. 2016.
- [25] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent Developments on ESPnet Toolkit Boosted by Conformer," *arXiv e-prints*, p. arXiv:2010.13956, Oct. 2020.
- [26] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech 2016*, 2016, pp. 2751–2755.
- [27] A. Stolcke, "Srlm — an extensible language modeling toolkit," *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, vol. 2, 07 2004.
- [28] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," *arXiv e-prints*, p. arXiv:1808.06226, Aug. 2018.
- [29] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," *arXiv e-prints*, p. arXiv:1804.00015, Mar. 2018.
- [30] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.
- [31] A. Rouzi, S. YIN, Z. ZHANG, D. WANG, H. Askar, and F. ZHENG, "Thuyg-20: A free uyghur speech database," *Journal of Tsinghua University(Science and Technology)*, vol. 57, no. 2, p. 182, 2017.