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

The TNT Team System Descriptions of Cantonese and Mongolian for IARPA OpenASR20

Jing Zhao[*†], *Zhiqiang Lv*[#†], *Ambyera Han*[#], *Guan-Bo Wang*[*], *Guixin Shi*[*],
Jian Kang[#], *Jinghao Yan*[#], *Pengfei Hu*[#], *Shen Huang*[#1$] *and Wei-Qiang Zhang*[*2$]

[*]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[#]TEG AI, Tencent Inc, Beijing 100193, China
[1]springhuang@tencent.com, [2]wqzhang@tsinghua.edu.cn

Abstract

This paper presents our work for OpenASR20 Challenge. We describe our Automatic Speech Recognition (ASR) systems for Cantonese and Mongolian under both constrained and unconstrained conditions. For constrained condition, a hybrid NN-HMM ASR system play the main role, while for unconstrained condition, an end-to-end ASR system outperforms traditional hybrid systems significantly due to adequate training data. Besides, we adapt to the challenging PSTN conditions using publicly available wideband dictated speech with similar accent, respectively for the two languages. Furthermore, data cleanup, language tailored features, multi-band training, data augmentation, pre-training and system fusions are incorporated. Our submitted systems have achieved excellent performances for the two conditions.

Index Terms: automatic speech recognition, low resource languages, OpenASR20, speech pretraining

1. Introduction

Due to the lack of speech data, language script, lexicons, building an applicable ASR system for low resource language is challenging. The goal of the OpenASR20 Challenge is to assess the state-of-the-art ASR technologies under low-resource language constraints. It consists of performing ASR on audio datasets in up to ten different low-resource languages, producing the recognized written text. For constrained condition, participants are only given 10 hours subset of labelled acoustic data but extra text data is unlimited. The datasets for most of the languages stem from the IARPA Babel program. For unconstrained condition, teams may use speech data in addition to the 10-hour subset provided for the language being processed, as well as additional publicly available speech and text training data from any other languages. The evaluation dataset is provided a week before the system submission deadline. We sign up for Cantonese and Mongolian in both constrained and unconstrained conditions. We describe our ASR systems in detail with necessary post analyses under constrained and unconstrained conditions in Sec.2 and Sec.3 respectively.

2. Constrained System

Since NN-HMM hybrid acoustic model is proved to be more promising in terms of performances for ASR than end-to-end (e2e) structures in the particular under-resource condition [1], hybrid structure is employed throughout the constrained condition. For the hybrid acoustic model, we propose the CNN-TDNN-F-A network as the main structure, which is trained with lattice-free maximum mutual information (LF-MMI) criterion [2]. The model introduces self-attention mechanism [3] to the combination of CNN and TDNN-F [4] in order to learn more positional information from the input. Specially, we adopt self-training strategy on the front-end of the hybrid system to obtain more effective representations with only 10h speech data. For low-resource condition, various data augmentations are combined to obtain additive improvement, such as speed and volume perturbation [5], Spec-Augment [6], Wav-Augment [7], adding noise as well as reverberation [8]. These are proved to be effective to ASR performance especially under low-resource conditions. Besides, we have trained more than four systems for each language to make further use of the diversities of different single systems by system fusion.

2.1. Speech Pretraining

Recently, speech pretraining has achieved giant improvement with large amount of unlabeled speech data [9]. Since extra speech data is not allowed for constraint condition, we try to excavate more effective representations for the limited 10 hours speech with pretraining.

The Transformer Encoder Representation for Speech (Tera) is adopted for feature extraction, which is composed of 3-layer Transformer Encoder layers [9]. We trained a Tera model with the constrained speech data to extract the 768-dimension representations. The hybrid systems trained with Tera-based representations converge much faster than MFCC acoustic features, though they do not outperform traditional features due to the lack of speech data as showed in Tab. 1. However, it is fortunate that Tera-based systems provide quite effective compensation for traditional features at system fusion phase.

2.2. Acoustic Model and Language Model

We adopt CNN-TDNN-F-A architecture as the main acoustic model, which combines Convolution Neural Network (CNN), Factored Time Delay Neural Network (TDNN-F) [4] and Self-Attention Network (SAN) [3, 10]. The CNN-TDNN-F-A network contains 11 TDNN-F blocks (9 before SAN and 2 after) in total with hidden dimension of 768 and a bottleneck dimension of 160. We adopt 6 convolution blocks at the beginning with concatenation of i-vectors and hires-MFCCs as input. Besides, other neural network acoustic models such as TDNN-F, TDNN-F-BLSTM(p), DFSMN [11], are also explored as additional fused systems. For training procedure, the acoustic model is trained with chain model in Kaldi [12] using LF-MMI criterion with cross-entropy (CE) regularization [2]. Apart from MFCC, for Cantonese 3-dim pitch features are added addition-

[†]equal contribution
[$]corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-1063

ally, which are more compatible with tonal languages. For speaker-aware training, i-vectors are trained based on a diagonal UBM derived extractor [13].

N-gram language models are applied in the hybrid system trained by SRILM [14], with extra text data from "training" part of IARPA Babel program[1] in addition to the provided data. Specifically, "IARPA-babel101b-v0.4c-build" is utilized for Cantonese, and "IARPA-babel401b-v2.0b-build" is used for Mongolian. Besides, we also employ lattice rescoring with NN-based LMs, which are composed by several TDNN-LSTM networks [15]. Two stages of LM rescoring using original and reversed text are adopted. In each stage a heuristic score for each arc to be expanded in the output lattice is computed to decide whether an arc is kept considering both historical and future information in the lattice [16].

Table 1: *Results for MFCC and Tera representations of Cantonese (DEV set, scoring locally)*

Features	WER	CER
Hires-MFCC	0.487	0.456
Hires-MFCC+Pitch	0.485	0.444
Tera representations	0.510	0.476
1+2+3 fusion	**0.468**	**0.424**

2.3. Data Processing

Removing noises is essential for data quality while adding diversities in the trained models is critical for data robustness.

2.3.1. Clean-up

Data clean-up is performed to remove the bad portions of the training transcripts with a biased ASR which is built by N-gram biased language model with garbage states projecting the nuisance parts of the speech. The operation seems to be effective for Cantonese.

2.3.2. Augmentation

In order to make full use of the limited training data, we adopt several popular techniques at the same time to enhance the data robustness and make our system more invariant to properties of the evaluation data.

- Speed and Volume Perturbation [5];
- Reverb and Noise Perturbation [7]: Noise from MUSAN database is chosen as stationary noise and noise from other speakers of the training corpus is chosen as babble noise. Reverberation is applied using the simulated RIRs with room sizes uniformly sampled from 1 to 30 meters;
- Wav-Augment [7]: We adopt pitch modification to deal with pitch changes. Band rejective filtering and time masking are employed to enhance the robustness of frequency and time domains, respectively;
- Spec-Augment [6]: Warping the features, masking blocks of frequency channels, and masking blocks of time steps;

The alignments for these augmented data are obtained from their clean counterparts. All the approaches are randomly chosen to enrich speech data in 20x. We do not make apple-to-apple comparisons among the four methods. However, we empirically summarize three rules: 1) Do not apply multiple augmentations for a single utterance; 2) Do not overuse data simulation as convergence of results will be observed at some point; 3) Speed perturbation is the most complementary approach that can be safely utilized.

2.4. Results

Our systems' performance under constrained condition is shown in Tab.2. The results for EVAL and that on the right-hand side of the arrow for DEV are released by OpenASR20 scoring server[2], while the left-hand side ones are obtained locally. For the evaluation period, Speech Activity Detection (SAD) is adopted following work in [17]. Besides, for EVAL lattice fusion [18] is applied first after decoding, which fuses hypothesis in more paths from resulting graph that yields better results than 1-best, followed by system fusion with ROVER [19]. Since the results on DEV set are from first-pass systems without any other pre-post processing, there's no wonder why the results on EVAL are better than DEV.

Table 2: *WER of ASR systems under constrained condition (For Mongolian, the results are submitted after close time)*

Features	DEV	EVAL
Cantonese	0.483 → 0.412	0.402
Mongolian	0.524 → 0.461	0.449

3. Unconstrained System

For unconstrained condition, e2e training, hybrid bandwidth acoustic modeling, language optimization and hybrid-e2e fusions are explored additionally.

3.1. Speech Pretraining

Since it is allowed to use extra publicly available data for training acoustic model under unconstrained condition. All the available 25 IARPA babel language training corpus are utilized to train the Tera pretrained model mentioned before. From preliminary results, the WER of systems with pretraining representations is nearly 10% lower than custom features relatively.

3.2. End-to-End System

It is shown that e2e ASR system strikes better performance than hybrid system in rich resource conditions and has high compatibility to legacy system as well [20]. We adopt Conformer structure [21] which is composed of two parts: encoder and decoder. The main body for the encoder is a convolution subsampling layer and several conformer blocks, which can be divided into four modules: a feed-forward module, a self-attention module, a convolution module, and a second feed-forward module in the end. The decoder part in Conformer is just the same as in Transformer. A simplified version of Conformer block, consisting of Relative Positional Embedding (RPE) introduced in

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

[2] https://www.nist.gov/itl/iad/mig/openasr20-challenge-results

Transformer-XL [22], CNN and Feedforward NN, is proposed without WER degradation, which is implemented in ESPnet [23]. In order to deal with long utterances and data sparsity in low-resource telephony condition, some techniques are proposed in the system:

- RPE [22]: The strategy can deal with repetition and instability brought by Transformer in long utterances;
- FL: Focal Loss is helpful to cope with unbalanced distribution of tokens;
- Ss: Scheduled Sampling is introduced to solve the problem of inconsistency during training and inference;

In our preliminary experiments, the Conformer based system achieves 3% to 10% relative improvements over traditional Transformer (0.45→0.43) in Cantonese. The number of Conformer blocks in encoder and decoder is 12 and 6, respectively. The encoder and decoder dimension are both 2048. The attention layer contains 4 heads and 256 units per head.

3.3. Acoustic Model for Hybrid System

For acoustic model, we have accumulated additional training data for Cantonese and Mongolian. Unfortunately, most of the data are wideband 16kHz speeches while the provided speech is recorded in telephony channel with a sampling rate of 8kHz. In order to utilize these data, we first train a mixed bandwidth acoustic model [24] with non-overlapping set of band-wise filters in 0-4kHz and 4-8kHz as illustrated in Fig. 1. For 8kHz speech, Spectral Band Replication (SBR) [25] is involved both in training and testing to fill out high frequencies, resulting in a feature extractor for 16kHz.

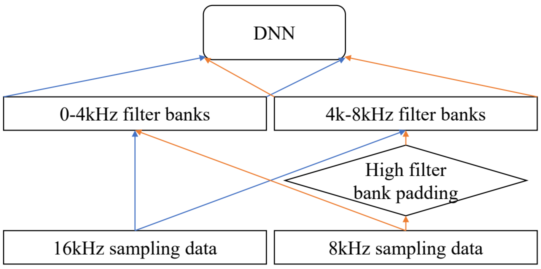

Figure 1: *Multi-band architecture for 8-16kHz hybrid speech recognition.*

For Cantonese, 2000h wideband dictated speeches from Speech Ocean Inc. and Huiting Tech Inc. and 140h narrow band (8kHz) speeches from IARPA Babel are applied. For Cyrillic Mongolian, only 10h wideband speeches from Mozilla and 50h narrowband speech from IARPA Babel are available. However, we believe that rich resources of Inner Mongolian speech in China is also beneficial to Cyrillic Mongolian speech recognition. An extra 500h wideband dictated speech from Speech Ocean and 100h dictated speech from M2ASR project [26] are incorporated. The goal is to utilize automatic Traditional to Cyrillic Mongolian conversion to make them compatible with each other. Details are illustrated in Tab. 3.

All the acoustic models are trained with lexicon from IARPA Babel program. Three-stream system is proposed:1) The first-way system is trained with original 8kHz Babel data by hybrid model; 2) For the second-way system, all the 8kHz and 16kHz speeches are trained with the above mixed bandwidth training strategy, after which the model is fine-tuned on SBR 16kHz OpenASR20 training data of the target language; 3) For the third-way system, the above mixed wideband e2e model is trained and tuned towards the same SBR 16kHz OpenASR20 data. For the first two hybrid systems mentioned above, the results are obtained by combining various acoustic models. The whole procedures for the main streams of system are illustrated in Fig. 2. The first two hybrid systems output lattices that can be rescored and fused using lattice combination [18]. The final lattice is timed aligned to form a CTM[6] result, which is then fused with e2e CTM via ROVER on 1-best sequence level.

Table 3: *Extra datasets for Cantonese and Mongolian*

Language	Narrowband 8kHz		Wideband 16kHz	
	source	**Duration**	**source**	**Duration**
Cantonese	Babel	140h	Speech Ocean[3]	1000h
	-	-	Huiting Tech[4]	1000h
Mongolian	Babel	50h	Mozilla[5]	10h
Inner Mongolian	-	-	M2ASR	100h
	-	-	Speech Ocean	500h

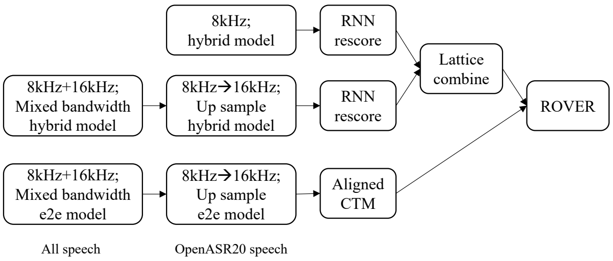
8kHz; hybrid model
RNN rescore
Lattice combine
8kHz+16kHz; Mixed bandwidth hybrid model
8kHz→16kHz; Up sample hybrid model
RNN rescore
ROVER
8kHz+16kHz; Mixed bandwidth e2e model
8kHz→16kHz; Up sample e2e model
Aligned CTM
All speech
OpenASR20 speech

Figure 2: *3-way multi-bandwidth hybrid & e2e system fusions.*

3.4. Language Optimization

It is easy to acquire large amount of text data using crawling engine from public available news. However, the domain mismatch is a severe problem for extra language model as performance on DEV set is deteriorated considerably with extra crawled text. Therefore, the crawled text of Cantonese (about 2GB) is only used for word segmentation to reduce WER (not CER). The main setup is the same with constrained condition, except that speech transcripts corresponding to the above extra acoustic data are incorporated in the training resources. Since the Cantonese vernacular text data is irregular, we use regular methods to correct common errors in the text data, such as abbreviations and typos after web crawling. For e2e system, characters are used as modeling unit rather than words, which means text segmentation is necessary to obtain better WER. We

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

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

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

[6] a concatenation of time mark records for each word in each channel of a waveform

use all the crawled and Babel text data to train a text segmentation model by Cantonese BERT pre-training [27] as showed in Fig. 3. Our UER tool for multi-lingual Bert model[7] is applied to train a multi-lingual encoder from mono-lingual data of Mandarin and Cantonese; Then a word-segmentation model is trained with Mandarin segmented data from the above pre-trained BERT model; Finally, the word segmentation model is fine-tuned with Babel Cantonese dataset.

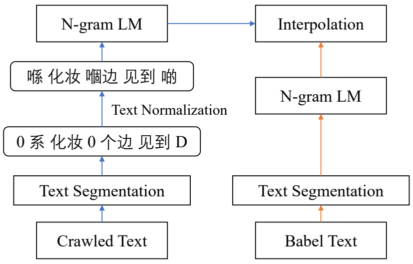

Figure 3: *Training word segmentation model (Cantonese).*

As known, there are several Mongolic languages or dialects which are roughly intelligible with each other. The speech data in OpenASR20 Mongolian is *Halh Mongolian* collected in Mongolia. Here we apply 500h dictated *Chakhar Mongolian* speech data provided by Speech Ocean, which is spoken in the Inner Mongolia region of China and sampled at 16kHz. Since *Halh Mongolian* and *Chakhar Mongolian* have some differences in the text (illustrated in Fig. 4), phoneme types and pronunciations. For acoustic model, we train the base model using Inner Mongolian speech and conduct transfer learning using the perturbed IARPA Babel *Halh Mongolian* speech data. The experiments show that transfer learning reduces the WER by 15% relatively. For language model, a seq2seq transformer model is trained to transfer *traditional Mongolian* text to *Mongolian Cyrillic* text to enrich *Cyrillic* language model.

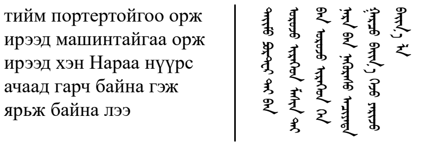
тийм портертойгоо орж
ирээд машинтайгаа орж
ирээд хэн Нараа нүүрс
ачаад гарч байна гэж
ярьж байна лээ

Figure 4: *Left: the Mongolian Cyrillic text used in Mongolia; Right: Traditional Mongolian text used in the Inner Mongolia region of China.*

3.5. System Fusion

For better performance, we build several different systems for system fusion as showed in Tab. 4. The Cantonese results of all the fused systems in DEV set are as below, system s1 and s2 are results from hybrid models, whereas s1 is trained by 8kHz Babel data, s2 is trained by extra 16kHz data with mixed bandwidth training, which shows that extra data and high frequency bring WER from 0.431 to 0.410. Meanwhile, e2e ASR system outperforms hybrid Model in single system with WER of 0.386. Using lattice combination, fused system s1 and s2 strikes WER of 0.404 against single system, meaning that unconstrained 16kHz data has a good compensation with Babel 8kHz data. The final result fused by s1, s2 and s3 with ROVER achieves best performance from 0.386 to 0.370. With extra post confidence filtering, WER can be further reduced to 0.361.

Table 4: *Unconstrained System Fusion Results for Cantonese (DEV set, scoring locally)*

System	Data&Model	Bandwidth	WER	CER
s1	8k Babel; chain	8k	0.431	0.398
s2	8+16k all; chain	Mixed band	0.410	0.370
s3	8+16k all; e2e	Mixed band	0.386	0.347
s1;s2	Fusion	-	0.404	0.370
s1;s2;3	Fusion	-	0.370	0.343
s1;s2;s3 filter	Fusion	-	0.361	0.332

3.6. Results on EVAL Set

Our systems' performance of unconstrained condition is shown in Tab. 5, which are released by OpenASR20 scoring server[2]. The results are much better than our local scoring results in Tab. 4 because we count language miscues, pauses, and other non-verbal speech as errors.

Table 5: *WER of ASR Systems (Unconstrained)*

Language	DEV	EVAL
Cantonese	0.335	0.320
Mongolian	0.381	0.406*

It can be observed that by using extra data, an absolute 7-8% WER reduction can be achieved in Cantonese and the CER is even much lower than 0.300 (0.264). In practice, speech recognition accuracy for Sino-Tibetan languages relies much on CER rather than WER, since WER for the languages is largely dominated by word segmentation errors, which may incur a biased result. We didn't manage to submit the fused system of Mongolian of unconstrained condition, so 0.406 is just a single HMM-NN hybrid system performance.

4. Conclusions

We achieve promising ASR results for Cantonese and Mongolian under constrained and unconstrained conditions with a series of fine but general techniques in the IARPA OpenASR20 challenge. The paper introduces the whole processes in detail to provide references for future research.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.U1836219 as well as research fund of Tsinghua University-Tencent Joint Laboratory for Internet Innovation Technology. And special thanks are given to all who supported TNT team during this challenge.

[7] https://github.com/dbiir/UER-py

6. References

[1] Y. Wang, D. Snyder, H. Xu *et al.*, "The JHU ASR system for VOiCES from a distance challenge 2019," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 2488–2492.

[2] D. Povey, V. Peddinti, D. Galvez *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, San Francisco, USA, Sep. 2016, pp. 2751–2755.

[3] D. Povey, H. Hadian, P. Ghahremani *et al.*, "A time-restricted self-attention layer for ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, Canada, Apr. 2018, pp. 5874–5878.

[4] D. Povey, G. Cheng, Y. Wang *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 3743–3747.

[5] T. Ko, V. Peddinti, D. Povey *et al.*, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3586–3589.

[6] D. S. Park, W. Chan, Y. Zhang *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 2613–2617.

[7] E. Kharitonov, M. Rivière, G. Synnaeve *et al.*, "Data augmenting contrastive learning of speech representations in the time domain," in *Proc. IEEE Spoken Language Technol. Workshop*, Shenzhen, China, Dec. 2020, pp. 1–5.

[8] T. Ko, V. Peddinti, D. Povey *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, New Orleans, USA, Mar. 2017, pp. 5220–5224.

[9] A. T. Liu, S.-W. Li, and H.-Y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," arXiv:2007.06028, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, USA, Dec. 2017, pp. 5998–6008.

[11] S. Zhang, C. Liu, H. Jiang *et al.*, "Nonrecurrent neural structure for long-term dependence," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 871–884, Feb. 2017.

[12] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, Waikoloa, USA, Dec. 2011, pp. 1–4.

[13] G. Saon, H. Soltau, D. Nahamoo *et al.*, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, Olomouc, Czech Republic, Dec. 2013, pp. 55–59.

[14] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. INTERSPEECH*, Denver, USA, Sep. 2002, pp. 901–904.

[15] X. Liu, X. Chen, Y. Wang *et al.*, "Efficient lattice rescoring using recurrent neural network language models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 8, pp. 1438–1449, Apr. 2016.

[16] H. Xu, T. Chen, D. Gao *et al.*, "A pruned RNNLM lattice-rescoring algorithm for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, Canada, Apr. 2018, pp. 5929–5933.

[17] G.-B. Wang and W.-Q. Zhang, "A fusion model for robust voice activity detection," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Ajman, United Arab Emirates, Dec. 2019, pp. 1–5.

[18] H. Xu, D. Povey, L. Mangu *et al.*, "An improved consensus-like method for minimum Bayes risk decoding and lattice combination," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Dallas, USA, Mar. 2010, pp. 4938–4941.

[19] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, Santa Barbara, USA, Dec. 1997, pp. 347–354.

[20] J. H. Wong, Y. Gaur, R. Zhao *et al.*, "Combination of end-to-end and hybrid models for speech recognition," in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 1783–1787.

[21] A. Gulati, J. Qin, C.-C. Chiu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 1783–1787.

[22] Z. Dai, Z. Yang, Y. Yang *et al.*, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. Annual Meeting Assoc. Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 2978–2988.

[23] S. Watanabe, T. Hori, S. Karita *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 2207–2211.

[24] J. Li, D. Yu, J.-T. Huang *et al.*, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. IEEE Spoken Language Technol. Workshop*, Miami, USA, Dec. 2012, pp. 131–136.

[25] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. IEEE Benelux Workshop Model Based Process. Coding Audio*, Leuven, Belgium, Nov. 2002, pp. 53–58.

[26] D. Wang, T. F. Zheng, Z. Tang *et al.*, "M2ASR: Ambitions and first year progress," in *Proc. Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, Seoul, Korea, Nov. 2017, pp. 208–213.

[27] J. Devlin, M.-W. Chang, K. Lee *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Computational Linguistics*, Minneapolis, USA, Jun. 2019, pp. 4171–4186.