ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094690

Multi-modal ASR error correction with joint ASR error detection

Binghuai Lin[1], *Liyuan Wang*[1]

[1]Smart Platform Product Department, Tencent Technology Co., Ltd, China

{binghuailin, sumerlywang}@tencent.com

Abstract

To tackle the recognition error problem for Automatic speech recognition (ASR), one common approach is to utilize text-based ASR error correction methods focusing on text error patterns. To include the audio information for better error correction, we propose a sequence-to-sequence multi-modal ASR error correction model. The multi-modal representations from pre-trained audio and text encoders are fused and aligned based on an attention mechanism. The decoder then searches for the corresponding correction results based on the fused representations. To better explore the correlations between different modalities, an additional ASR error detection task is applied on top of the fused representations. We optimize the network by a multi-task learning method combining both ASR error detection and correction tasks. Experimental results based on a 200-hour dataset recorded by Chinese English-as-second-language (ESL) learners show the proposed correction model can achieve significant improvement compared to the baselines with or without other ASR error correction methods.

Index Terms: ASR error correction, attention, pre-trained, multi-modal, error detection

1. Introduction

With the significant improvement in accuracy of automatic speech recognition (ASR), it has been widely utilized in various speech-based applications, such as speech translation and spoken dialogue. Despite the advances in ASR, there are still unavoidable recognition errors due to multiple factors, including different domains, accents, distant ASR, etc. [1], which hamper the broader application of ASR.

End-to-end (E2E) ASR systems have prevailed in recent studies. Instead of training separate models including the acoustic model (AM), the pronunciation model (PM), and the language model (LM) in traditional speech recognition systems, the E2E ASR systems optimize them jointly using a single network [2][3][4]. As E2E ASR systems only use paired speech-text data for training, leaving a large amount of text data unused [5], many studies have proposed to improve the ASR performance further with the additional text data. One common solution is through ASR error correction.

Many ASR error correction methods have been proposed based on text-only data, including ASR recognition results and transcriptions. Previous works can be classified into two categories : (1) neural language models (LM) with fusion techniques; and (2) error correction models mapping recognition results with errors to ground truth sentences. LM fusion commonly trains an external LM to be incorporated into the ASR system, and based on the fusion mechanism, it can be further categorized into different approaches. Shallow fusion combined the predicted probability of the decoder and the external LM during inference [6]. Deep fusion concatenated the output hidden states of the attention-based system and the pre-trained LM, and then fine-tuned the combining parameters [7]. Cold fusion trained the E2E systems from scratch with a fixed pre-trained LM model [8]. Though these methods have proved to be effective, there are still problems with rare words and proper nouns, which could be due to the fact that the integration of LM and ASR doesn't aim at correcting ASR errors [9]. Error correction models investigate the error patterns of ASR models by mapping recognition results with errors to ground-truth sentences. Some works treated the detection and correction of ASR errors as two separate steps [10]. Other works directly corrected ASR errors without extra ASR error detection. An encoder-decoder structure was utilized to rescore the generative probability of words to correct recognition errors based on the N-best lists and ASR confidence scores generated in a speech recognizer [11]. With the development of neural networks (NN), many end-to-end architectures have been proposed for ASR error correction. For the CTC-based speech recognition system, a transformer-based spelling correction model was proposed with the recognition results as input and the ground-truth transcriptions as output [12][13]. To correct spelling errors in recognition results, error hypotheses generated by the Listen, Attend and Spell (LAS) model [4] and the corresponding ground truth was utilized for training a sequence-to-sequence spelling correction model [9]. Machine translation model was proposed for ASR error correction, and it has been proved that the downstream task of speaker diarization improved with the improvement of the stylistic characteristics of transcription [14]. Transformer with BERT and copying mechanism was proposed to further improve ASR error correction performance [1].

For better performance of ASR error correction, some work utilizes both speech and ASR hypotheses as the input instead of text-only data. A deliberation network was proposed to encode both acoustics and first-pass hypotheses and concatenate these context vectors into a LAS decoder [15]. To alleviate the overfitting problem of the text-only correction model, acoustic features called Mel filter-bank are concatenated with the text features and fed to the spelling correction model [5]. Inputs of different modalities, including speech and text, are jointly encoded based on a cross-modal self-attention mechanism in the ASR error correction model to capture the correlations between them [16]. A cross-modal ASR error correction model was proposed to combine error correction, and utterance rejection tasks, which was proved to be effective for both single-speaker and multi-speaker speech [17]. Inspired by these studies, we propose a multi-modal ASR error correction method aiming at better fusing modalities of speech and text. Soft alignment between text and acoustic representations are achieved based on the attention mechanism, and the fusion is explicitly guided by an auxiliary ASR error detection task, presumably to explore the correlations between different modalities to facilitate the correction task.

These authors annotated with [1] contributed equally to this work.

Furthermore, for better representations of the text and speech modalities, we exploit the strengths of pre-trained models prevailing in speech and natural language processing (NLP) [18][19]. Downstream tasks can achieve satisfying performance by fine-tuning with only a few labeled data. In this paper, we propose a multi-modal ASR error correction model combining acoustic and text representations, which are derived from pre-trained wav2vec 2.0 and BERT models [18][19]. The error correction and the auxiliary error detection tasks are optimized by a multi-task learning (MTL) method. Experimental results on the dataset recorded by Chinese English-as-second-language (ESL) learners show the proposed method can achieve performance improvement compared to the ASR baseline and other ASR error correction methods. In section 2, we will introduce the proposed method. The experiments are conducted in section 3. We will draw the conclusion and future suggestions in section 4.

2. Proposed method

The proposed network is composed of an acoustic encoder, a text encoder, and a text decoder, as shown in Figure 1. The acoustic encoder takes speech signals as input and outputs acoustic representations. The text encoder generates text representations of ASR hypotheses derived from an E2E ASR model as input. These two kinds of representations are fused by an attention mechanism. The attention-weighted acoustic representations and text representations are added and fed into a text decoder to obtain the corrected texts. An auxiliary ASR error detection is proposed to assist with fusing these modalities, and the whole network is optimized by MTL.

2.1. E2E ASR model

Our E2E ASR model takes raw speech signal as input and generates hypotheses as output. It shares a similar structure to the speech transformer, which consists of an acoustic encoder and a text decoder [20]. In our work, the acoustic encoder is based on wav2vec 2.0 [19], which is a framework for self-supervised learning of representations from raw audio data. It comprises a multi-layer convolutional feature encoder and a context network, which follows the Transformer architecture [21], optimized by contrastive loss based on large amounts of unlabeled data. The decoder is based on the BERT structure [18]. The encoder takes frame acoustic features as input and outputs the contextual representations of acoustic features. The decoder generates the final output based on the final hidden state derived from a stack of self-attention and cross-attention-based modules conditioned on the contextual representations from the encoder. The model is trained by minimizing the cross-entropy loss on the training data.

2.2. Multi-modal ASR error correction model

The multi-modal ASR error correction extracts acoustic and text representations by an acoustic encoder and a text encoder.

The acoustic representations are denoted as $\mathrm{H}_{\text{speech}} = \{h^1_{\text{speech}}, h^2_{\text{speech}}, \cdots, h^m_{\text{speech}}\}$, where h^i_{speech} is the acoustic feature for the ith frame. These representations are extracted from a pre-trained wav2vec 2.0 acoustic encoder [19].

The text encoder and decoder are BERT-based structures taking ASR hypotheses generated from the E2E ASR model as input [18]. The text representations derived from the text encoder are denoted as $\mathrm{H}_{\text{text}} = \{h^1_{\text{text}}, h^2_{\text{text}}, \cdots, h^n_{\text{text}}\}$, where h^j_{text} is the representation for the jth subword tokenized by a Sentencepiece tokenizer [22].

The representations from different modals are fused through a multi-head attention (MHA) mechanism. The attention can be defined as Eq. (1):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where Q denotes the queries, and K denotes the keys with the dimension of d_k. V represents the values with the dimension of d_v. In our paper, we take the representations of jth subwords h^j_{text} as queries and representations of speech $\mathrm{H}_{\text{speech}}$ as keys and values. The final jth attention-weighted acoustic representation is shown as Eq. (2):

$$\left(h^j_{\text{text}}\right)' = \text{Attention}(h^j_{\text{text}}, \mathrm{H}_{\text{speech}}, \mathrm{H}_{\text{speech}}) \tag{2}$$

The weighed acoustic representations and the raw subword representations are fused by addition as shown in Eq. (3):

$$h_j = \left(h^j_{\text{text}}\right)' + h^j_{\text{text}} \tag{3}$$

Finally, the decoder takes the previous subword and the fused contextual representations as input and outputs the next subword.

2.3. Network optimization and inference

Attention-based fusion for different modalities has been proved effective in many areas. To better fuse these modalities automatically, it usually relies on sufficient training data or needs some appropriate auxiliary tasks. Due to the data scarcity problem in ASR error correction, we pre-train the attention-based fusion module first and then optimize the network with the auxiliary ASR error detection task.

First, we pre-train the encoders and attention-based fusion module with a synthetic dataset from E2E ASR training data, consisting of paired speech-text data. Specifically, we randomly replace the text words of the paired data with the rest words in the CMU pronunciation dictionary [23] with the probability of replacement to be 40%. The replaced words are labeled as 1, and the rest remain labeled as 0. We add an extra fully connected layer following the multi-modal representations in Eq. (3) to predict whether the word is replaced. As the replacing words don't exist in the speech, we can treat the task as ASR error detection of errors in the hypotheses. The pre-training is optimized with cross-entropy loss as defined in Eq. (4), where n is the total number of the words in the hypothesis, y^i_{error} represents the ith word label (1 for replaced words and 0 for the rest), and p^i_{error} is the ith predicted probability indicating whether the word is replaced.

$$L_{\text{detec}} = -\sum_{i=1}^{n} y^i_{\text{error}} \times \log(p^i_{\text{error}}) \tag{4}$$

After pre-training, we fine-tune the pre-trained encoders and train the decoder with ASR error correction data, consisting of human-transcribed text, ASR-generated hypothesis, and speech utterance. The optimization predicts whether the word in a hypothesis sentence is erroneous and decodes the corresponding human-transcribed words for a particular speech utterance and hypothesis. The correction loss L_{cor} is defined by cross-entropy loss between the human-transcribed text y_{word} and the predicted sequence p_{word} as shown in Eq. (5), where k is the total number of the words in the human-transcribed text. The whole network is optimized in an MTL framework

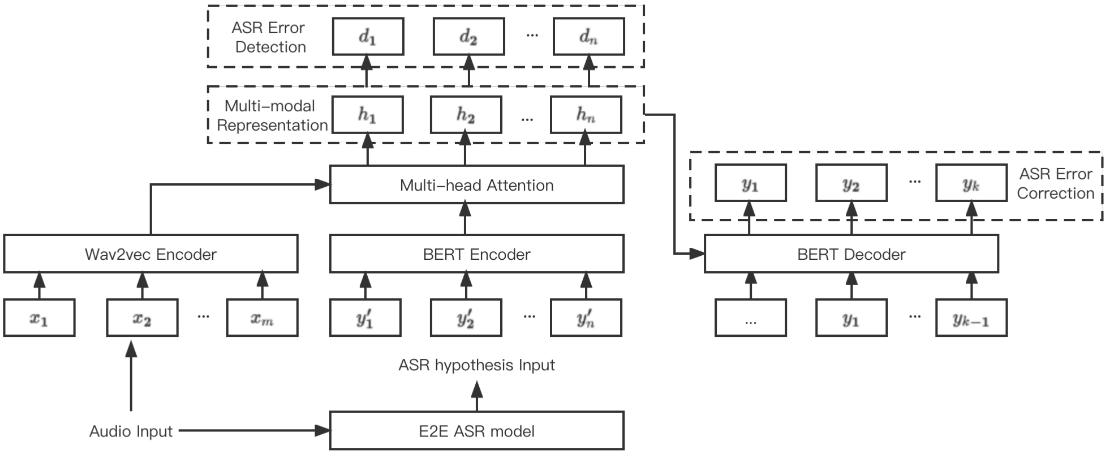

Figure 1: *Network structure of multi-modal ASR error correction*

combining the losses from the ASR error correction and detection as shown in Eq. (6), where α ranging from 0 to 1 is the hyper-parameter to balance these two losses.

$$L_{\text{cor}} = -\sum_{i=1}^{k} y^{i}_{\text{word}} \times \log(p^{i}_{\text{word}}) \tag{5}$$

$$L_{\text{total}} = L_{\text{cor}} + \alpha * L_{\text{detec}} \tag{6}$$

During inference, the encoder takes the raw speech and its ASR hypothesis as input, and the decoder attends to multi-modal representations from encoders and generates the corrected sequence by choosing the most likely symbol at each time step.

3. Experiments

3.1. Corpus

The baseline E2E ASR model is trained based on the 960-hour native LibriSpeech corpus [24]. For the ASR error correction model, the training data consist of a 200-hour non-native corpus recorded by 757 Chinese teenagers, containing 150,000 utterances with an average duration of 5 seconds. The ASR error correction data are transcribed by three annotators, with the final transcripts obtained by a majority vote. For pre-training error detection, we randomly choose 200-hour data from the LibriSpeech corpus and construct synthetic data as mentioned in Section 2.3. For fine-tuning, we construct correction data based on the non-native data, consisting of ASR hypotheses generated from the trained E2E ASR model and corresponding human-transcribed transcripts. We split the data into 70% for training, 10% for validation, and 20% for testing for ASR error correction.

3.2. Experimental setup

Our experiments are performed using Huggingface Transformers [25]. The encoder of the E2E ASR model is initialized from wav2vec2-base-960h [1], which is composed of a feature encoder containing seven temporal convolutions blocks and a 12-layer transformer encoder. The ASR decoder is initialized from the bert-base-uncased model [2], which is composed of a 12-layer transformer encoder. We trained the ASR model based on the LibriSpeech corpus and achieved a word error rate (WER) of 3.5% in the test-clean dataset and 8.5% in the test-other dataset. Based on the trained ASR model, we generated hypotheses for the non-native data with a WER of 21.3%.

The speech encoder of the ASR error correction model shares the same structure as the encoder of the ASR model, and the text encoder and decoder are initialized from the bert-base-uncased model. The cross attention module is one MHA layer fusing acoustic representations with the dimensionality of 768 and text representations with the same dimensionality derived from the corresponding speech and text encoders. The decoder predicts subwords of 30522 classes. We use Adam optimizer with $\beta 1 = 0.9, \beta 2 = 0.99$. The learning rate is set with a warm-up scheme with an initial learning rate of $5e-5$ and a warm-up ratio of 0.1. The α for MTL training is 0.5 in our experiments.

We compare our results with different multi-modal and single-modal baselines. For multi-modal baselines, we adopt the same fusion strategies similar to previous works, including the model taking the concatenated down-sampled raw acoustic features called Mel filter-bank and ASR hypotheses as input [5] (referred to as Mel+Text) and the model using cross-modal encoder to encode the concatenated acoustic and text embeddings [16] (referred to as Cross-Modal). We also experiment with a cascading method combining two separate tasks, including audio fine-tuning and ASR error correction (referred to as Audio→Text). For a fair comparison, we utilize the same structures of the text encoder and decoder as ours for these models.

For single-modal baselines, we compare our results with the LM fusion method (referred to as LM shallow fusion), text-only ASR error correction method (referred to as Text-only), and audio-only fine-tuning method (referred to as Audio-only). For the LM fusion method, we train an LM including two unidirectional LSTM layers with the dimensionality of 1024 based on the transcripts of our non-native training data. The ASR model and LM are combined by shallow fusion. The text-only error correction is composed of a BERT-based text encoder and decoder trained with the paired ASR hypotheses and human tran-

[1] https://huggingface.co/facebook/wav2vec2-base-960h

[2] https://huggingface.co/bert-base-uncased

scripts. The audio-only fine-tuning method is to fine-tune the E2E ASR model with our non-native data.

3.3. Experimental results

First, we show the comparison with different baselines. Then, we perform some ablation studies to demonstrate the effectiveness of the proposed method. Finally, we will make some analysis on the results of different baselines to show the superiority of the proposed method.

3.3.1. Comparison results

The results are shown in Table 1. As the ASR baseline is trained based on the LibriSpeech corpus, deviating from our non-native training data, it performs worst in WER. It can be found that models with single-modality perform worse than those based on multi-modalities, indicating superiority of multi-modal fusion for ASR error correction.

For single-modal baselines, Audio-only performs best, indicating the benefits of exploiting non-native speech. For multi-modal baselines, we can see that the method with the cross-modal encoder performs equally well with the cascading method. The proposed method performs best, which achieves 17.6% and 12.2% relative reduction in WER compared with single-modal baselines and multi-modal baselines, respectively.

Table 1: *WER (%) of single-modal and multi-modal baselines*

Modality	Model	WER
-	E2E ASR baseline	21.3
Single-Modal	LM shallow fusion	16.2
	Text-only	15.5
	Audio-only	13.1
Multi-Modal	Mel+Text [5]	12.8
	Cross-Modal[16]	12.3
	Audio→Text	12.5
	Ours	**10.8**

3.3.2. Ablation studies

To show the effectiveness of the proposed method, we carry out some ablation studies. First, as the proposed method is composed of two stages, including ASR error detection pre-training and MTL training, we compare to results without the pre-training stage. Second, we compare to results trained with pre-training, but single-task learning (STL) in the second stage. Results are shown in Table 2. From the results, we can see the pre-training stage has significant benefits for the proposed method. The MTL can further improve performance compared with the STL baseline.

Table 2: *WER (%) of different variants of the proposed model*

Model	WER
No pre-training	14.8
STL	12.4
Ours	**10.8**

3.3.3. Analysis on different baselines

We make some analysis on the results of different baselines. First, we will show the number of insertion, deletion, and substitution errors in the test dataset as shown in Figure 2. From the figure, the proposed method performs better than other baselines in reducing different kinds of errors, and it performs best in reducing substitution errors.

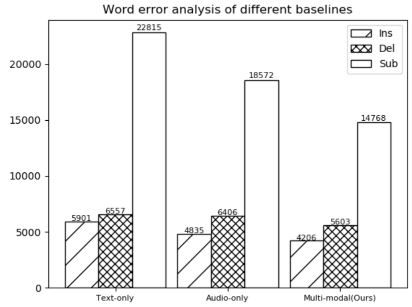

Figure 2: *Error analyses of different baselines*

Then, we conduct some qualitative analysis of some examples from the single-modal baselines and the proposed method. The examples are shown in Table 3. From the results, we can see that the ASR baseline makes errors on words with similar pronunciation, some proper nouns, or common phrases. The Text-only method can fix some errors of proper nouns and common phrases. For example, “it is” can be decoded into “it’s”, and “shan hai” can be converted into “shanghai”. The Audio-only can fix some words error with close pronunciation such as “play” and “plane”, and “close” and “clothes”. By combining two modalities, the proposed method can solve both kinds of errors.

Table 3: *Examples of different ASR error correction methods*

Model	Example 1
Ground-truth	Take the **plane** to **shanghai it’s** faster
E2E ASR	Take the **play** to **shan hai it is** faster
Text-only	Take the **place** to **shanghai it’s** faster
Audio-only	Take the **plane** to **shanghai it is** faster
Multi-modal(Ours)	Take the **plane** to **shanghai it’s** faster
Model	Example 2
Ground-truth	He **made some little clothes** for them
E2E ASR	He **make same it a close** for them
Text-only	He **made it a close** for them
Audio-only	He **made some good clothes** for them
Multi-modal(Ours)	He **made some little clothes** for them

4. Conclusion

This paper proposes a multi-modal ASR error correction method combining acoustic and text representations. We initialize the acoustic and text encoders with the pre-trained wav2vec 2.0 and BERT for better representation. For better fusion of these modalities, we utilize the multi-head attention mechanism and perform pre-training with an auxiliary ASR error detection task with synthetic data. The whole network is then fine-tuned by MTL based on the ASR error correction data. Experimental results based on the non-native data demonstrate the superiority of the proposed method, with 17.6% and 12.2% relative reduction in WER compared with single-modal and multi-modal baselines, respectively. It is worthwhile applying the proposed method to different downstream tasks such as spoken language assessment for further investigation in the future.

5. References

[1] W. Li, H. Di, L. Wang, K. Ouchi, and J. Lu, "Boost Transformer with BERT and copying mechanism for ASR error correction," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–6.

[2] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[3] 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)*. IEEE, 2018, pp. 5884–5888.

[4] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[5] S. Zhang, J. Yi, Z. Tian, Y. Bai, J. Tao, X. Liu, and Z. Wen, "End-to-End Spelling Correction Conditioned on Acoustic Feature for Code-Switching Speech Recognition," *Proc. Interspeech 2021*, pp. 266–270, 2021.

[6] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang, "Shallow-fusion end-to-end contextual biasing," in *Interspeech*, 2019, pp. 1418–1422.

[7] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[8] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[9] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5651–5655.

[10] Y. Bassil and P. Semaan, "Asr context-sensitive error correction based on microsoft n-gram dataset," *arXiv preprint arXiv:1203.5262*, 2012.

[11] T. Tanaka, R. Masumura, H. Masataki, and Y. Aono, "Neural Error Corrective Language Models for Automatic Speech Recognition," in *INTERSPEECH*, 2018, pp. 401–405.

[12] S. Zhang, M. Lei, and Z. Yan, "Automatic spelling correction with transformer for ctc-based end-to-end speech recognition," *arXiv preprint arXiv:1904.10045*, 2019.

[13] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of automatic speech recognition with transformer sequence-to-sequence model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7074–7078.

[14] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "Asr error correction and domain adaptation using machine translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6344–6348.

[15] K. Hu, T. N. Sainath, R. Pang, and R. Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7799–7803.

[16] T. Tanaka, R. Masumura, M. Ihori, A. Takashima, T. Moriya, T. Ashihara, S. Orihashi, and N. Makishima, "Cross-Modal Transformer-Based Neural Correction Models for Automatic Speech Recognition," *arXiv preprint arXiv:2107.01569*, 2021.

[17] J. Du, S. Pu, Q. Dong, C. Jin, X. Qi, D. Gu, R. Wu, and H. Zhou, "Cross-Modal ASR Post-Processing System for Error Correction and Utterance Rejection," *arXiv preprint arXiv:2201.03313*, 2022.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] 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)*. IEEE, 2018, pp. 5884–5888.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[23] R. L. Weide, "The CMU pronouncing dictionary," *URL: http://www.speech.cs.cmu.edu/cgibin/cmudict*, 1998.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.