TWO-STAGE PRE-TRAINING FOR SEQUENCE TO SEQUENCE SPEECH RECOGNITION

Zhiyun Fan[1,2], Shiyu Zhou[1], Bo Xu[1]
[1]*Institute of Automation, Chinese Academy of Sciences, China*
[2]*School of Artificial Intelligence, University of Chinese Academy of Sciences, China*
{fanzhiyun2017, zhoushiyu2013, xubo}@ia.ac.cn

***Abstract*—The attention-based encoder-decoder structure is popular in automatic speech recognition (ASR). However, it relies heavily on transcribed data. In this paper, we propose a novel pre-training strategy for the encoder-decoder sequence-to-sequence (seq2seq) model by utilizing unpaired speech and transcripts. The pre-training process consists of two stages, acoustic pre-training and linguistic pre-training. In the acoustic pre-training stage, we use a large amount of speech to pre-train the encoder by predicting masked speech feature chunks with their contexts. In the linguistic pre-training stage, we first generate synthesized speech from a large number of transcripts using a text-to-speech (TTS) system and then use the synthesized paired data to pre-train the decoder. The two-stage pre-training is conducted on the AISHELL-2 dataset, and we apply this pre-trained model to multiple subsets of AISHELL-1 and HKUST for post-training. As the size of the subset increases, we obtain relative character error rate reduction (CERR) from 38.24% to 7.88% on AISHELL-1 and from 12.00% to 1.20% on HKUST.**

***Index Terms*—pre-training, speech recognition, encoder-decoder, sequence-to-sequence**

I. Introduction

There have been growing interests in building an end-to-end (E2E) speech recognition system, which directly transforms an input sequence of acoustic features to an output sequence of tokens. The single all-neural structure makes the E2E system have several advantages, including a simpler training process and joint optimization among components. Currently prominent E2E models include: (a) connectionist temporal classification (CTC) [1], [2], (b) attention-based encoder-decoder networks [3]–[7], and (c) recurrent neural network transducer (RNN-T) [8].

Although E2E models are powerful, they still suffer from the problem that the training process is very hungry for human-transcripted supervised data. Unfortunately, the collection of supervised data is time-consuming and expensive. Comparing with supervised data, unpaired data (speech and text) is much easier to collect. One of the solutions to reduce the need for paired data is to use speech and text respectively. Thus a lot of unsupervised and semi-supervised methods [9]–[12] were proposed to ease the dependence. The success of these unsupervised and semi-supervised methods indicates that there is useful semantic knowledge in these unpaired speech and text. They can be used separately.

This work is supported by the Key Research and Development Program of the Ministry of Science and Technology under No. 2017YFB1002102.

Recently, the release of BERT (Bidirectional Encoder Representations from Transformers) [13] provides us a new way to utilize unpaired data by pre-training. BERT is a bidirectional variant of Transformer networks trained to jointly predict a masked word from its context and to classify whether two sentences are consecutive or not. Then the pre-trained model can swiftly adapt for downstream tasks by fine-tuning. It obtains new state-of-the-art results on eleven natural language processing tasks. According to [13], BERT can capture the structural information about language contained in text-only data by pre-training and this semantic information is helpful to downstream tasks. Intuitively, as another carrier of semantic information, speech can be processed similarly. Recent researches [14]–[16] used BERT-style pre-training for the ASR system. Jiang et al. [14] proposed Masked Predictive Coding (MPC), a method utilizing Masked-LM like structure for Transformer based speech recognition models. Wang et al. [15] pre-trained bidirectional RNNs for direct use in a CTC based speech recognizer and explored both time- and frequency-domain masking. Baevski et al. [16] proposed vq-wav2vec to learn discrete representations of audio segments before leveraging BERT-style pre-training. All these works only focus on how to pre-train the acoustic encoder. For the encoder-decoder framework, the pre-training of the decoder also needs consideration.

In this paper, we propose a two-stage pre-training for the attention-based encoder-decoder framework. Two pre-training stages are leveraged to extract acoustic and linguistic information from speech and transcripts respectively. In the first stage, we pre-train the encoder with a large amount of unlabeled speech data. We mask some continuous feature chunks in each sequence at random and use contexts to predict them. Using this BERT-style pre-training, we obtain good initial parameters for the encoder. In the second stage, we first generate the speech from a large number of transcripts with a trained text-to-speech (TTS) [17] system and then use these synthesized data to optimize the whole network. Although the acoustic information of synthesized data is monotonous, transcripts contain rich linguistic information, which is useful for the downstream ASR task.

All of our experiments are conducted on the Transformer [7], an encoder-decoder framework. After the two-stage pre-training with AISHELL-2, we fine-tune the pre-trained model on AISHELL-1 and HKUST during post-training. We use

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534170

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on March 01,2022 at 02:28:10 UTC from IEEE Xplore. Restrictions apply.

multiple subsets of AISHELL-1 or HKUST as paired data to conduct post-training. As the size of the subset increases, we obtain relative character error rate reduction (CERR) from 38.24% to 7.88% on the test set of AISHELL-1. For the HKUST dataset, we obtain relative CERR from 12.00% to 1.20%.

II. RELATED WORK

The most related work to this paper is BERT [7], which is a bidirectional language representation model. BERT is trained to capture useful representations by predicting masked tokens with their context and classifying the relationship between two sentences. Thus when fine-tuned on downstream tasks, the model can converge faster and better than initializing with scratch. As for the ASR task, similarly, we design the first pre-training stage to capture useful representations contained in the speech. Our pre-training policy is different from BERT mainly in two aspects: (i) We mask a continuous feature vector sequence rather than discrete tokens. Unlike text which can be broken into character or word units relatively easily, speech features are continuous, and the neighboring frames are similar. So only predicting discrete frames is too easy for neural networks. (ii) We discard the next sentence prediction mentioned in [7]. Because the relationships between sentences are not important to the ASR task.

Inspired by BERT, pre-training attracted increasing attention in the field of speech recognition [14]–[16]. Jiang et al. [14] verified the effect of BERT-style pre-training on the encoder of the Transformer. Using tens of thousands of hours of speech during pre-training, it achieved obvious improvement. In this paper, we also employ similar BERT-style pre-training for the encoder. Besides, we use synthesized paired data for pre-training the decoder in our linguistic pre-training. Wang et al. [15] investigated the effectiveness of pre-training on phone-based and character-based CTC systems. And it was the first to pre-train bidirectional RNNs for the speech recognizer. In this paper, we use the Transformer, a seq2seq framework, as the study platform and explore the influence caused by the size and distribution of data. Baevski et al. [16] pre-trained a feature extractor with discretized unlabeled speech data. The extractor is used to generate discretized speech representations, which are used to train the acoustic model.

III. TWO-STAGE PRE-TRAINING

In this paper, we use the Transformer as the study platform to investigate our two-stage pre-training method, which consists of acoustic pre-training and linguistic pre-training. And a post-training is employed to fine-tune the pre-trained model for the downstream ASR task. The entire training process is shown in Fig. 1. As an attention-based encoder-decoder model, the Transformer can be divided into three parts, that is, encoder, decoder, and cross-attention. The acoustic pre-training aims to integrate useful representations contained in speech into the encoder by predicting some masked feature chunks. In the linguistic pre-training, we use a trained TTS system to generate speech from a large number of transcripts. Using these synthesized paired data, the decoder can obtain rich linguistic information. During post-training, paired data is used to fine-tune the model. The details are as follows.

A. Acoustic pre-training

The left side of Fig. 1 illustrates the acoustic pre-training. To pre-train the encoder, we mask some continuous feature sequences of input $x = (x_1, x_2, ..., x_T)$ along the time steps, where T is the length of the input sequence. The width and position of these masked feature chunks are sampled randomly. We set the masked value in chunks to zero, and only predict these masked features rather than reconstruct the entire input. The details of our mask strategy are shown as follows:

- K masked chunks: Firstly, K time points are chosen along the time steps $(0, T)$ as centers of these feature chunks, denoted as c_i. And chunk width w is sampled from a uniform distribution from 0 to W. Thus $2W + 1$ is the max length that can be masked during training.

$$c_i \sim uniform(0, T), i \in (1, 2, ..., K) \tag{1}$$

$$w \sim uniform(0, W) \tag{2}$$

  We denote the time interval of the i-th masked feature chunk as f_i. The s_i and e_i are the start and end of the i-th chunk respectively.

$$s_i = max(0, c_i - w), e_i = min(c_i + w, T) \tag{3}$$

$$f_i = [s_i, e_i], i \in (1, 2, ..., K) \tag{4}$$

- 80% of the time: For each feature chunk, there is an 80% chance that values in the chunk are masked to zero.

$$x'_t = \mathbf{0}, t \in f_i \tag{5}$$

- 20% of the time: The feature chunks are not always masked. There is a 20% chance that the values in chunk stay the same. The purpose of this is to bias the representation towards the actual speech feature sequence.

The encoder reads a masked sequence of d-dimensional feature vector $x' = (x'_1, x'_2, ..., x'_T)$, and transforms it to higher-order representations $h = (h_1, h_2, ..., h_T)$. At the top of the encoder, an extra linear layer projects hidden features to the same dimension as input features. This final output sequence can be regarded as a hypothesis of input feature sequence, denoted as $\hat{x} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_T)$.

Instead of reconstructing the entire input, we only predict these masked features. We use mean square error (MSE) loss to conduct training.

$$Loss = \frac{1}{BK} \sum_{b=1}^{B} \sum_{i=1}^{K} \sum_{t \in f_i} \|x_{b,t} - \hat{x}_{b,t}\|^2 \tag{6}$$

where the subscript (b) indicates the b-th example in a batch which contains B examples.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on March 01,2022 at 02:28:10 UTC from IEEE Xplore. Restrictions apply.

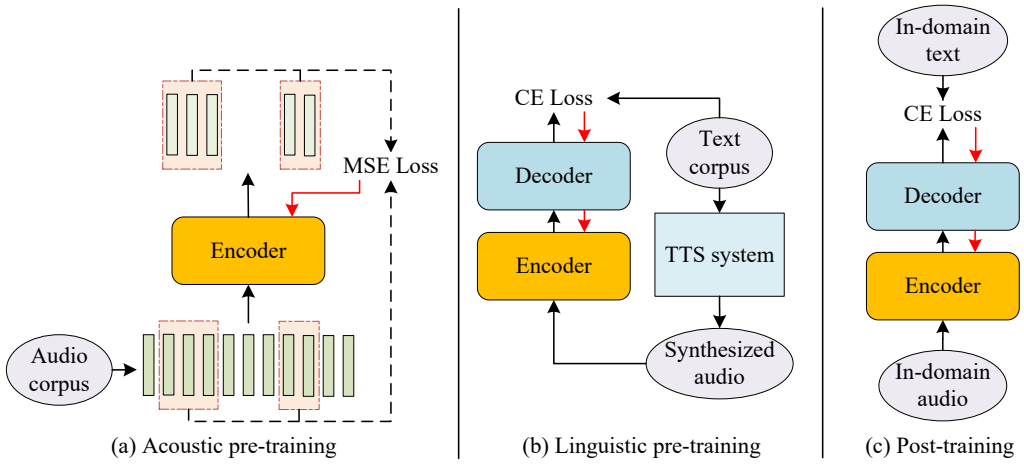

Fig. 1. Overview of our two-stage pre-training method. (a) Acoustic pre-training is applied by predicting masked feature chunks with contexts. (b) TTS system generates paired data from a large number of transcripts, which is used to conduct linguistic pre-training. (c) A schematic representation of post-training.

B. Linguistic pre-training

ASR is a speech-to-text task and model like the Transformer usually consists of an encoder and a decoder. Only using our acoustic pre-training proposed in section III-A to pre-train encoder is not enough for this encoder-decoder framework. This section proposes a linguistic pre-training to pre-train the decoder by using a large number of transcripts.

A common approach to leverage text-only data is training a language model (LM), and fusing the output of the decoder and the pre-trained LM. Deep fusion [18] and shallow fusion [19] are two ways to integrate LM into the E2E system. Although these fusion approaches have shown improvements to the E2E system, the drawback is that the extra LM increases complexity to the system. In this paper, we propose a linguistic pre-training to integrate linguistic information into the seq2seq system. Specifically, we use a trained TTS system to generate speech for a large number of transcripts, which converts text-only data to paired data. Then we use these synthesized paired data to train a Transformer whose encoder is initialized with acoustic pre-training mentioned in section III-A. This training stage uses cross-entropy (CE) loss, and backpropagation updates the whole model, including the encoder, the decoder, and the cross-attention. Although the acoustic information of these synthesized paired data is monotonous, the linguistic information contained in these transcripts is rich. This process is illustrated in the middle of Fig. 1.

Compared with extra LM [18], [19] and BERT [20], using synthesized paired data to integrate linguistic information has two advantages: (i) The decoder of Transformer stacks identity blocks which contain three sublayers, i.e., self-attention, encoder-decoder attention, and feed-forward network. BERT-initialized decoder can only initialize two sublayers, and the encoder-decoder attention is still initialized randomly. However, the encoder-decoder attention represents the alignment between speech and text. Fortunately, our method which is training with synthesized paired data can help the system to capture both linguistic information and alignment between speech and text. (ii) The training stage with synthesized paired data integrates linguistic information into the decoder without extra LM, which leads to a more simple model structure and reduces the complexity of the system.

C. Post-training

Using our two-stage pre-training, the seq2seq model extracts rich acoustic and linguistic representations that are useful for the downstream ASR task. After the pre-training, an extra post-training is necessary. The pre-trained model is fine-tuned with paired data during the post-training. The right side of Fig. 1 illustrates this process. In this stage, supervised training is conducted by CE loss, and the model is initialized with the last checkpoints in linguistic pre-training. The softmax layer is reinitialized randomly, and the number of output units depends on the training set. During fine-tuning, we update the parameters of encoder, decoder, and cross-attention simultaneously.

IV. EXPERIMENTS

A. Datasets

We experiment on three public ASR datasets including AISHELL-2 [21], AISHELL-1 [22], and HKUST [23]. AISHELL-2, a Mandarin ASR dataset, contains about 1000 hours of speech-to-text data. AISHELL-1, a subset of AISHELL-2, contains about 178 hours of speech. HKUST is a spontaneous speech corpus (201 hours), whose recording environment and language style are quite different from AISHELL-2. During the two-stage pre-training, speech and transcripts of AISHELL-2 are used to conduct the acoustic and linguistic pre-training respectively. Due to the inclusion relationship between AISHELL-1 and AISHELL-2, we remove these sentences appearing in the test set of AISHELL-1 from AISHELL-2. AISHELL-1 and HKUST are used to fine-tune the model in the post-training stage. The difference is that the distribution of AISHELL-1 is consistent with AISHELL-2, but

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on March 01,2022 at 02:28:10 UTC from IEEE Xplore. Restrictions apply.

HKUST is not. HKUST is used to simulate the scenario that the distribution of post-training data and pre-training data is inconsistent.

The synthesized data used in the linguistic pre-training is generated from 99392 transcripts (remove 7176 transcripts in the test set of AISHELL-1) in AISHELL-2, and the duration of synthesized audio is up to 800 hours. The synthesis system is a Tacotron2 trained with an open high-quality Mandarin Chinese dataset which consists of about 12 hours of speech data [24]. The structure and training details of the speech synthesis system can be found in [17].

B. Modeling and training

All the acoustic features used in this paper are 80-dimensional log-Mel filter-bank features, computed with a 25 ms window and shifted every 10 ms. The raw features are normalized via mean subtraction and variance normalization per speaker. Before flowing into the model, the features are firstly stacked with 3 frames to the left and then down-sampled to 33.3 Hz frame rate. Because the length of the speech feature sequence varies from tens to thousands, we mask the stacked feature sequence with $K = 2$ and $W = 10$. Efforts on adjusting the two hyperparameters can bring little improvement. Thus this paper does not discuss how to set K and W.

Transformer used in this paper contains 6 encoder-blocks and 6 decoder-blocks, with a per-block configuration of $d_{model} = 512$, attention heads $h = 16$, and feed-forward inner-layer dimension $d_{ff} = 2048$. In the linguistic pre-training, we use 3961 characters appearing in the AISHELL-2 and 4 extra tokens, including an unknown token (<UNK>), a padding token (<PAD>), and sentence start and end tokens (<S>/<\S>) as output units. In the post-training, the softmax layer is reinitialized randomly. We use 4230 and 3896 characters plus 4 extra tokens as output units for AISHELL-1 and HKUST respectively.

During both pre-training and post-training, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\varepsilon = 10^{-9}$ and alter the learning rate over the course of training according to the formula:

$$lrate = k \cdot d_{model}^{-0.5} \cdot \min(n^{-0.5}, n \cdot warmup\_n^{-1.5}) \quad (7)$$

where n is the step number, k is a tunable scalar and the learning rate increases linearly for the first $warmup\_n$ training steps and decreases thereafter proportionally to the inverse square root of the step number. The $warmup\_n$ steps are 12000 and 8000 for linguistic pre-training and post-training respectively.

In the linguistic pre-training and post-training, the label smoothing of value $\varepsilon_{ls} = 0.1$ is employed [25]. And the last 20 checkpoints are averaged for inference. For evaluation, we use beam search with a beam size of 13 and length penalty $\alpha = 0.6$.

C. Pre-training models

The acoustic pre-training only uses speech of AISHELL-2 to train the encoder of the Transformer, which is stacked an extra full connected (FC) layer that projects dimension of features to 320. After the acoustic pre-training stage, we discard the extra FC layer and only keep the parameters of the encoder, denoted as M-a. In linguistic pre-training, we use the transcripts of AISHELL-2 to train a whole Transformer whose encoder is initialized with M-a. We denote this model as M-al. To conduct the ablation study in section V-B, we apply our linguistic pre-training to a Transformer initialized randomly and denote this model as M-l. In addition, we pre-train a Transformer initialized randomly with real paired AISHELL-2, denoted as M-sup. In summary, M-a, M-al, M-l, and M-sup represent acoustic pre-training model, linguistic pre-training model, two-stage pre-training model and supervised pre-training model respectively. All these models and their corresponding pre-training methods are listed in Table I.

TABLE I
INSTRUCTIONS OF ALL PRE-TRAINING MODELS USED IN THIS PAPER.

	M-a	M-al	M-l	M-sup
acoustic pre-training	✓	✓	×	×
linguistic pre-training	×	✓	✓	×
supervised pre-training	×	×	×	✓

V. RESULTS

A. Results on AISHELL-1

In this section, we evaluate our approach on AISHELL-1, whose distribution is consistent with AISHELL-2. We use 10 hours, 20 hours, 89 hours and 178 hours of AISHELL-1 as paired data for post-training respectively. Table II summarizes the CER on the test set of AISHELL-1. A0 is the baseline model initialized from scratch. The result of 7.87% (178 hours) shows that our baseline Transformer is competitive. Initialized with the two-stage pre-trained model (M-al), A1 obtains a relative CERR from 38.24% to 7.88% than the baseline system (A0) as the size of subsets of AISHELL-1 increase. It indicates that the two-stage pre-training benefits the downstream ASR task, and the greater the ratio of unpaired data to paired data, the better the effectiveness of the pre-training method. When this ratio reaches about one hundred to one (the case of 10h in the A1), the relative CERR can reach 38.24%, which means our pre-training method is more helpful to the low-resource scenarios. Besides, models initialized with the two-stage pre-training method converge consistently faster than randomly initialized baseline. Fig. 2 illustrates the loss

TABLE II
CER[%] PERFORMANCE OF THE TWO-STAGE PRE-TRAINING METHOD ON AISHELL-1.

Exp	Initial Models	10h	20h	89h	178h
A0	scratch	32.77	21.22	11.04	7.87
A1	M-al	**20.24**	**15.03**	**8.98**	**7.25**
A2	M-a	25.20	16.59	9.19	7.45
A3	M-l	23.31	16.70	9.43	7.56
A4	M-sup	18.08	13.12	8.16	6.70

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on March 01,2022 at 02:28:10 UTC from IEEE Xplore. Restrictions apply.

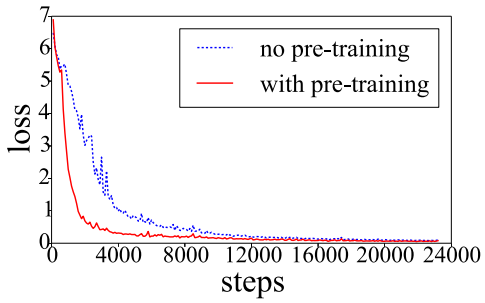

Fig. 2. Loss curve of post-training with or without pre-training when all 178 hours of AISHELL-1 are used.

curve of 178 hours case. The dotted line represents the baseline and the solid line is initialized with a pre-trained model. In another three cases, the loss curves show a similar tendency. Experiments A4 are initialized with a supervised pre-training model M-sup. A0 and A4 can be regarded as lower and upper bounds of our pre-training method. Comparing A1 with A4, we can find that the two-stage pre-training has been pretty close to the supervised pre-training.

B. Ablation study on AISHELL-1

In this section, we use several ablation studies to demonstrate the effectiveness of acoustic and linguistic pre-training respectively. A2 and A3 in Table II ablate linguistic pre-training and acoustic pre-training respectively. From the results is Table II, several observations can be found:

1) A2 vs. A0: All the experiments in A2 are initialized from acoustic pre-training model M-a. As the size of the subsets of AISHELL-1 increases, A2 obtains relative CERR from 23.10% to 5.34%. It means that acoustic pre-training can benefit downstream ASR tasks without the help of linguistic pre-training. And it shows a similar performance tendency as that in the two-stage pre-training.
2) A3 vs. A0: All the experiments in A3 are initialized from linguistic pre-training model M-l. A3 obtains relative CERR from 28.87% to 3.94% with only linguistic pre-training. It means that linguistic pre-training can also work independently of acoustic pre-training.
3) A2 vs. A3: When the amount of available paired data for the ASR task is very small (10 hours), the linguistic pre-training is more useful than acoustic pre-training. With the amount of paired data used in post-training increasing, the acoustic pre-training starts to play a more important role.
4) A2 and A3 vs. A1: All the experiments in A1 are initialized from two-stage pre-training. A1 obtains more relative CERR than A2 and A3. It indicates that combining acoustic and linguistic pre-training can further improve the performance and these two pre-training stages are complementary.

C. Results on HKUST

The distribution of AISHELL-1 and AISHELL-2 is similar. But in some scenarios, the distribution of paired data used in post-training is very different from the data used in the pre-training procedure. In this section, we evaluate our approach on one quarter, two quarters, three quarters and all of the HKUST dataset respectively. HKUST is a dataset that is very different from AISHELL-2 on the sampling rate, recording environment and language style. AISHELL-2 is a read-speech dataset, but HKUST is a spontaneous and informal speaking dataset.

TABLE III
CER[%] PERFORMANCE OF OUR UNSUPERVISED PRE-TRAINING METHOD ON HKUST.

Exp	Initial Models	1/4	1/2	3/4	ALL
B0	scratch	43.09	33.46	30.18	26.56
B1	M-a	**37.92**	**31.80**	29.49	**26.24**
B2	M-al	38.05	32.07	**29.28**	26.32
B3	M-l	40.29	33.16	30.10	26.51

Next, we discuss whether this mismatch will weaken the effectiveness of our method or not. In Table III, B0 is the results of our baseline Transformer on HKUST without pre-training. Comparing B1 with B0, we obtain relative CERR from 12.00% to 1.20%, which indicates that the acoustic pre-training boosts the downstream ASR task even if there is a mismatch between unpaired pre-training data and post-training data. In other words, the acoustic pre-training is robust to the change in sampling rate and recording environment. Comparing B3 with B0, we can find that linguistic pre-training almost has no benefits to the downstream ASR task when the language style of data used in pre-training and post-training is very different. It is intuitive. HKUST contains a lot of colloquial expressions, many of which do not even conform to the grammar. However, the transcripts of AISHELL-2 are more grammatical. The potential language models for the two datasets are very different. So the linguistic pre-training with AISHELL-2 can bring little improvement to HKUST. Furtherly, B2 applies two-stage pre-training to HKUST. Comparing B2 with B1, we find that in three cases B1 obtains better results. Only in the three quarters case, B2 is a little better than B1. This means that in such a mismatch scenario, linguistic pre-training brings almost no improvement even if it is used with acoustic pre-training together. So, the choice of speech used in acoustic pre-training is more free, and the distribution of the transcripts used in linguistic pre-training is best to be consistent with the paired data used in post-training.

VI. CONCLUSION

In this paper, we investigate the usage of unpaired speech and transcripts to conduct a two-stage pre-training. Experiments demonstrate these models, which leverage large-scale two-stage pre-training, outperform those that only use paired data. We obtain relative CERR from 38.24% to 7.88% on the test set of AISHELL-1 as the paired training set increases. In other words, the greater the ratio of unpaired data to paired data, the better the effectiveness of our pre-training method. Besides, we verify the effectiveness of our approach when there is a mismatch between pre-training and post-training

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on March 01,2022 at 02:28:10 UTC from IEEE Xplore. Restrictions apply.

data. We obtain relative CERR from 12.00% to 1.20% on HKUST as the paired training set increases. It means that our pre-training method is robust to data mismatches. And the consistency of data used in pre-training and post-training can bring more improvements. In the future, we plan to apply our approach to larger datasets and investigate more efficient unsupervised pre-training methods.

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. ACM, 2006, pp. 369–376.

[2] A.Graves, and N.Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in International conference on machine learning, 2014, pp. 1764–1772.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[5] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recog- nition," in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016, pp. 4945–4949.

[6] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," Advances in neural information processing systems, 2015, pp. 577–585.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[8] A. Graves, "Sequence transduction with recurrent neural net- works," arXiv preprint arXiv:1211.3711, 2012.

[9] K.-Y. Chen, C.-P. Tsai, D.-R. Liu, H.-Y. Lee, and L.-s. Lee, "Completely unsupervised phoneme recognition by a generative adversarial network harmonized with iteratively refined hidden markov models," arXiv preprint arXiv:1904.04100, 2019.

[10] Y.-A. Chung, W.-H. Weng, S. Tong and J. Glass, "Unsupervised cross-modal alignment of speech and text embedding spaces," in Advances in Neural Information Processing Systems, 2018, pp. 7354–7364.

[11] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, and M. Delcroix, "Semi-supervised end-to-end speech recognition," in Interspeech, 2018, pp. 2–6.

[12] B. Li, T. N. Sainath, R. Pang and Z. Wu, "Semi-supervised training for end-to-end models via weak distillation" in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2837–2841.

[13] G. Jawahar, B. Sagot, and D. Seddah, "What does bert learn about the structure of language?" in 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019.

[14] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving transformer-based speech recognition using unsupervised pre-training," arXiv preprint arXiv:1910.09932, 2019.

[15] W. Wang, Q. Tang, and K. Livescu, "Unsupervised pre-training of bidirectional speech encoders via masked reconstruction," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6889–6893.

[16] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," arXiv preprint arXiv:1910.05453, 2019.

[17] Y. Z. Zou, L. H. Dong, and B. Xu, "Boosting character-based chinese speech synthesis via multi-task learning and dictionary tutoring," Proc. Interspeech 2019, pp. 2055–2059, 2019.

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

[19] J. Chorowski, and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," arXiv preprint arXiv:1612.02695, 2016.

[20] 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.

[21] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," arXiv preprint arXiv:1808.10583, 2018.

[22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE, 2017, pp. 1–5.

[23] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "Hkust/mts: A very large scale mandarin telephone speech corpus," in International Symposium on Chinese Spoken Language Processing. Springer, 2006, pp. 724–735.

[24] "Chinese standard mandarin speech copus (10000 sentences)," http://www.data-baker.com/open source.html.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on March 01,2022 at 02:28:10 UTC from IEEE Xplore. Restrictions apply.