HIERARCHICAL CROSS-MODALITY KNOWLEDGE TRANSFER WITH SINKHORN ATTENTION FOR CTC-BASED ASR

Xugang Lu[1*], *Peng Shen*[1], *Yu Tsao*[2], *Hisashi Kawai*[1]

1. National Institute of Information and Communications Technology, Japan
2. Research Center for Information Technology Innovation, Academia Sinica, Taiwan

ABSTRACT

Due to the modality discrepancy between textual and acoustic modeling, efficiently transferring linguistic knowledge from a pretrained language model (PLM) to acoustic encoding for automatic speech recognition (ASR) still remains a challenging task. In this study, we propose a cross-modality knowledge transfer (CMKT) learning framework in a temporal connectionist temporal classification (CTC) based ASR system where hierarchical acoustic alignments with the linguistic representation are applied. Additionally, we propose the use of Sinkhorn attention in cross-modality alignment process, where the transformer attention is a special case of this Sinkhorn attention process. The CMKT learning is supposed to compel the acoustic encoder to encode rich linguistic knowledge for ASR. On the AISHELL-1 dataset, with CTC greedy decoding for inference (without using any language model), we achieved state-of-the-art performance with 3.64% and 3.94% character error rates (CERs) for the development and test sets, which corresponding to relative improvements of 34.18% and 34.88% compared to the baseline CTC-ASR system, respectively.

Index Terms— Pretrained language model (PLM), Cross-modality alignment, sinkhorn attention, automatic speech recognition (ASR)

1. INTRODUCTION

Due to the non-autoregressive (NAR) decoding capability for fast and parallel inference, the temporal connectionist temporal classification (CTC)-based learning [1] for automatic speech recognition (ASR) is one of the most attractive frameworks for end to end (E2E) ASR [2]. However, token independence assumption in CTC based learning makes it difficult for acoustic encoder to learn rich context dependent linguistic information. Leveraging a language model (LM), particularly a pretrained language models (PLM) to improve the ASR performance is a promising direction. In early studies [3], based on attention with encoder-decoder (AED) modeling, a hybrid CTC/AED-based ASR model framework was proposed to enhance linguistic information in acoustic encoder. With multi-task learning framework, several methods have been proposed to learn linguistic information by inserting linguistic knowledge in intermediate layers of acoustic encoders for ASR [4, 5]. In recent years, due to the success of self-supervised learning in feature exploration, knowledge transfer learning from both pretrained acoustic model (e.g., wav2vec2.0 [6]) and PLM (e.g., bidirectional encoder representation from transformers (BERT) [7]) for ASR also have been proposed [8, 9, 10, 11, 12].

Although stacking text encoder of a PLM on top of the acoustic encoder could improve the ASR performance[13], it is preferred to transfer linguistic knowledge encoded in the PLM to acoustic encoding via cross-modal knowledge distillation (KD) [11, 14, 15, 16, 17]. However, in most studies, the KD learning is carried out on the probability logits of the acoustic model or on the last hidden layer of acoustic encoders [11, 15, 14] . This learning paradigm is based on a simple assumption that high abstract-level of acoustic feature corresponds to tokens of linguistic knowledge in a bottom-up based processing of speech. However, as studies revealed that the acoustic feature learning even in low-level should be guided with linguistic knowledge as a top-down attention process [18]. In this study, we propose a novel cross-modality knowledge transfer (CMKT) learning framework for linguistic knowledge transfer from a PLM to acoustic encoding in CTC based ASR. Our main contributions are summarized as: 1. A hierarchical acoustic alignment with the linguistic latent representations from a PLM are applied for CMKT, even in low-level of acoustic features which is different from previous studies in [12, 19]; 2. In cross-modality alignment, we propose to use the Sinkhorn attention [20, 21] for feature alignment where the transformer attention [22] is a special case of the iteration of Sinkhorn normalization process; 3. We implemented the CMKT learning algorithm in a CTC based ASR for acoustic encoding and confirmed its effectiveness via detailed experiments.

2. PROPOSED METHOD

The proposed model framework is illustrated in Fig. 1, and the adapter and cross-modality matching modules in Fig. 1 are further explained in Fig. 2. In these figures, 'FC1', 'FC2', 'FC3' are linear transforms of fully-connected layers, 'LN'

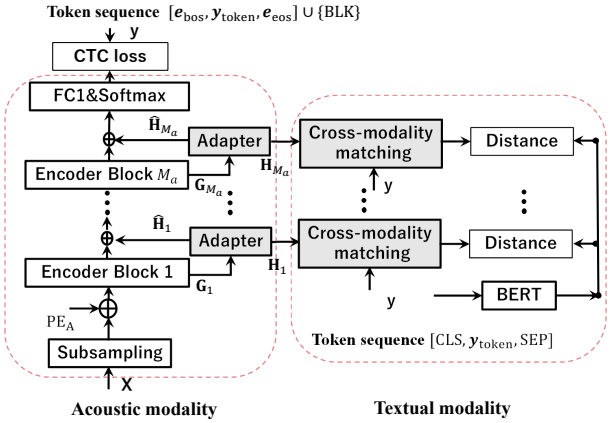

Fig. 1. The proposed cross-modality knowledge transfer framework for CTC-based ASR.

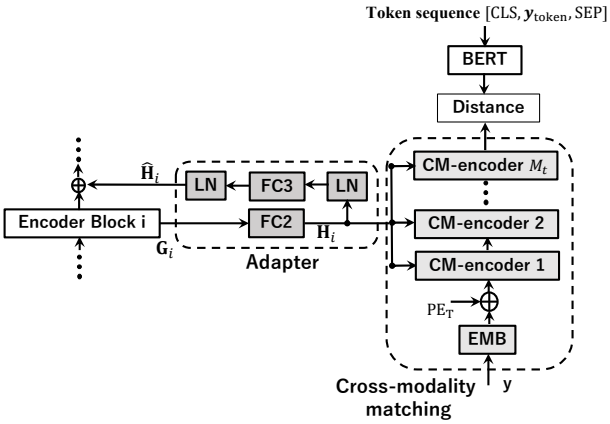

Fig. 2. Adapter and Cross-modality matching modules which are shared by all encoder blocks.

denotes layer-normalization, and 'CM-encoder' represents cross-modality encoder. In the follows, we will explain the process of each module.

2.1. Features from acoustic and textual modalities

In acoustic modality of Fig. 1, the process in 'Subsampling' module and position encoding of acoustic sequence ($\mathrm{PE_A}$) are used to extract the initial input feature as $\mathbf{G}_0$. Then the output from the i-th acoustic encoder block is represented as:

$$\mathbf{G}_i = \mathrm{Encoder}_i\left(\mathbf{G}_{i-1}\right) \in \mathbb{R}^{l_a \times d_a} \tag{1}$$

where i takes values from 1 to M_a, with M_a representing the total number of encoder blocks, l_a and d_a are length (temporal dimension) and feature dimension, respectively. In adapter process, a linear transform FC2 is applied for feature dimension matching between acoustic and textual modalities (in Fig. 2):

$$\mathbf{H}_i = \mathrm{FC}_2\left(\mathbf{G}_i\right) \in R^{l_a \times d_t} \tag{2}$$

In this $\mathbf{H}_i$, the feature dimension is d_t corresponding to textual feature dimension, and i is the encoder block index.

The initial textual feature is obtained from token embedding 'EMB' and position encoding of text '$\mathrm{PE_T}$' as $\mathbf{Z}_0 \in R^{l_t \times d_t}$ with sequence length l_t and feature dimension d_t. This $\mathbf{Z}_0$, together with representations from acoustic modality is transformed by a sequence of cross-modality encoders (as CM-encoder in Fig. 2). For an acoustic representation $\mathbf{H}$ from Eq. (2) (encoder block index is omitted for easy explanation), the transform in each CM-encoder is formulated as:

$$\mathbf{Z}_j = f_j(\mathbf{Z}_{j-1}, \mathbf{H}), \tag{3}$$

where j is the index of textual encoder block with values from 1 to M_t. In each CM-encoder, there is a sequential process with modality feature transform, layer-normalization, and feed forward transform as:

$$\mathbf{Z}_{\mathbf{H}\to\mathbf{Z}_{j-1}} = \mathrm{OT}\left(\mathbf{H} \to \mathbf{Z}_{j-1}\right) \tag{4}$$

$$\begin{aligned} \hat{\mathbf{Z}}_{j-1} &= \mathrm{LN}\left(\mathbf{Z}_{j-1} + \mathbf{Z}_{\mathbf{H}\to\mathbf{Z}_{j-1}}\right) \\ \mathbf{Z}_j &= \mathrm{LN}\left(\hat{\mathbf{Z}}_{j-1} + \mathrm{FC}\left(\hat{\mathbf{Z}}_{j-1}\right)\right) \end{aligned} \tag{5}$$

In Eq. (4), $\mathbf{Z}_{\mathbf{H}\to\mathbf{Z}_{j-1}}$ is a transported representation (from acoustic modality to textual modality), and $\mathrm{OT}(.)$ denotes optimal transport (OT). As it is showed that the acoustic feature $\mathbf{H} \in \mathbb{R}^{l_a \times d_a}$, and the textual feature $\mathbf{Z}_{j-1} \in \mathbb{R}^{l_t \times d_t}$ are two modalities with different lengths. After OT, the representation $\mathbf{Z}_{\mathbf{H}\to\mathbf{Z}_{j-1}}$ keeps the same dimensions of that of $\mathbf{Z}_{j-1}$. The final output of the cross-modality matching is represented as $\mathbf{Z}_{M_t}$. For transferring linguistic knowledge from the BERT, we suppose that this $\mathbf{Z}_{M_t}$ should approximate a target textual representation which is provided by the pretrained BERT model as:

$$\begin{aligned} &\mathbf{y}_{\mathrm{token}} = \mathrm{Tokenizer}\left(\mathbf{y}\right); \tilde{\mathbf{Z}}_0 = [\mathrm{CLS}, \mathbf{y}_{\mathrm{token}}, \mathrm{SEP}] \\ &\tilde{\mathbf{Z}}_m = \mathrm{BERT}_m\left(\tilde{\mathbf{Z}}_{m-1}\right) \in \mathbb{R}^{l_t \times d_t} \end{aligned} \tag{6}$$

where 'BERT_m' is the m-th transformer encoder layer of BERT model, m takes values from 1 to M_b, with M_b representing the total number of BERT encoder layers. 'Tokenizer' is a process to convert standard text to word piece based tokens [7]. Token symbols 'CLS' and 'SEP' represent the start and end of an input sequence. In model learning stage, for linguistic knowledge transfer, the loss function is defined as cross-modal alignment loss by:

$$L_{\mathrm{align}} = \sum_{j=2}^{l_t-1} 1 - \cos\left(\mathbf{z}_{j,:}, \tilde{\mathbf{z}}_{j,:}\right), \tag{7}$$

where $\mathbf{z}_{j,:}$ and $\tilde{\mathbf{z}}_{j,:}$ are row vectors of feature matrices $\mathbf{Z}_{M_t}$ and $\tilde{\mathbf{Z}}_i$ from 'BERT_i', respectively. In this formulation, the sum ranges from 2 to $l_t - 1$ in order to exclude the '[CLS]' and '[SEP]' tokens from the loss estimation (refer to Eq. (6) in text encoding).

2.2. Sinkhorn attention for CMKT learning

In Eq. (4), the solution is represented as (index subscript is omitted for easy explanation):

$$\mathbf{Z}_{\mathbf{H}\to\mathbf{Z}} = \hat{\gamma} \times \mathbf{H} \in R^{l_t \times d_t} \tag{8}$$

where $\hat{\gamma}$ is a transport coupling matrix based on minimizing an entropy regularized OT (EOT) $L_{EOT}(\mathbf{H}, \mathbf{Z})$ as:

$$\hat{\gamma} = \underset{\gamma \in \prod(\mathbf{H},\mathbf{Z})}{\arg\min}\ L_{EOT}(\mathbf{H}, \mathbf{Z}) \tag{9}$$

And the objective function $L_{EOT}(\mathbf{H}, \mathbf{Z})$ is defined as:

$$L_{EOT}(\mathbf{H}, \mathbf{Z}) \triangleq \sum_{i,j} \gamma_{i,j} C_{i,j} + \alpha \gamma_{i,j} \log \gamma_{i,j}, \tag{10}$$

where α is a regularization coefficient, $\gamma_{i,j}$ and $C_{i,j}$ are elements of transport coupling γ and cost matrices C. The solution of Eq. (9) can be implemented as an iteration of Sinkhorn projections as [20, 23]:

$$\gamma^0 = \exp\left(-\frac{1}{\alpha}C\right), \gamma^{k+1} = F_c\left(F_r\left(\gamma^k\right)\right), \tag{11}$$

where $F_c(.)$ and $F_r(.)$ are column- and row-wise normalization operators, respectively. In real applications, in Eq. (11), only a few iterations are enough to obtain fairly well results (in our experiments, 3 times of iterations were set).

In iteration process of Sinkhorn attention, the row-wise normalization $F_r(.)$ can be formulated as:

$$F_r(\gamma) = \frac{\gamma}{\sum_j \gamma_{i,j}} = \mathrm{softmax}\left(\mathrm{FC}\left(C\right)\right), \tag{12}$$

where 'FC' is a linear transform. When choosing negative inner product as the cost function for C, Eq. (12) is further cast to:

$$F_r(\gamma) = \mathrm{softmax}\left(\mathbf{Z}\mathbf{W}_{\mathbf{Z}}\left(\mathbf{H}\mathbf{W}_{\mathbf{H}}\right)^T\right), \tag{13}$$

where $\mathbf{W}_{\mathbf{Z}}$ and $\mathbf{W}_{\mathbf{H}}$ are feature transform matrices for acoustic and textual representations, respectively. From this equation, we can see that the transformer attention [22] can be regarded as a special case of the Sinkhorn attention with proper chosen of linear transforms and cost functions which also have been studied in [21].

2.3. Loss function in CMKT learning

For transferring back linguistic information in acoustic encoding, the following transforms are designed as indicated in Fig. 2:

$$\begin{aligned} \hat{\mathbf{H}}_i &= \mathrm{FC3}\left(\mathrm{LN}\left(\mathbf{H}_i\right)\right) \in R^{l_a \times d_a} \\ \mathbf{H}_i^{a,t} &= \mathbf{G}_i + \mathrm{LN}\left(\hat{\mathbf{H}}_i\right) \end{aligned} \tag{14}$$

Based on this new representation $\mathbf{H}^{a,t}$ which is supposed to encode both acoustic and linguistic information, the final probability prediction for ASR is formulated as:

$$\tilde{\mathbf{P}} = \mathrm{Softmax}\left(\mathrm{FC1}\left(\mathbf{H}_{M_a}^{a,t}\right)\right) \tag{15}$$

In training with CMKT, the total loss is defined as:

$$L \triangleq \lambda . L_{\mathrm{CTC}}(\tilde{\mathbf{P}}, \mathbf{y}_{\mathrm{token}}) + (1-\lambda) . w . \sum_{i=1}^{M_a} \left(L_{\mathrm{align}}^i + L_{\mathrm{EOT}}^i\right), \tag{16}$$

where $L_{\mathrm{CTC}}(\tilde{\mathbf{P}}, \mathbf{y}_{\mathrm{token}})$ is CTC loss, L_{align}^i and L_{EOT}^i are cross-modality alignment loss and OT loss collected from the hierarchical encoder block indexed by i as defined in Eqs. (7) and (10), respectively. λ is a trade off parameter, w is a parameter to scale the alignment loss. After the model is trained, only the left branch of Fig. 1 is kept for ASR inference.

3. EXPERIMENTS

We carried out experiments on an open source Mandarin speech corpus AISHELL-1 which includes speech recorded from 400 speakers [24]. Three data sets are included: a training set with 340 speakers (150 hours), a development (or validation) set with 40 speakers (10 hours), and a test set with 20 speakers (5 hours). In training, data augmentation with speed perturbation (with factors of 0.9 and 1.1) was applied [24]. 80-dimensional log Mel-filter bank features together with 3-dimensional fundamental frequency related features (F0, delta F0 and delta delta F0) are used as raw input feature, and they were extracted with a 25ms window size and a 10ms shift.

3.1. Parameter settings

In acoustic modality, the convolutional block in CNN subsampling module is with 256 channels, kernel size 3, stride 2, and ReLU activation function. Conformer based acoustic encoder [25] is used. In each conformer block, the convolution is with kernel size of 15, attention dimension is $d_a = 256$, attention head is 4, and the dimension of FFN layer is 2048. The BERT of 'bert-base-chinese' from huggingface is used as the PLM. In this BERT model, there are $M_b = 12$ transformer encoders, token size is 21128, and text feature dimension is $d_t = 768$. For reducing calculation redundancy, the CMTK was carried out on every three layers of the acoustic encoders. Several other hyper-parameters are fixed and set as: EOT regularization parameter $\alpha = 1.0$ in Eqs. (10) and (11), scale parameter $w = 1.0$ and alignment trade off parameter $\lambda = 0.3$ in Eq. (16). In optimization, Adam optimizer [26] is used with a learning rate (initial with 0.001) schedule with 20,000 warm-up steps. The model was trained for 130 epochs, and the final model used for evaluation was obtained by averaging models from the last 10 epochs. The performance was evaluated based on character error rate (CER).

Table 1. ASR performance on AISHELL-1 coprus, CER (%).

Methods	dev set	test set
Conformer-CTC (Baseline)	5.53	6.05
Conformer-CTC/AED ([3])	4.61	5.06
NAR-BERT-ASR ([13])	4.90	5.50
LASO with BERT ([11])	5.20	5.80
KT-RL-ATT ([9])	4.38	4.73
Wav2vec-BERT ([12])	4.10	4.39
Last-CMKT (proposed)	**4.05**	**4.40**
Hierarchical-CMKT (proposed)	**3.64**	**3.94**

3.2. Results

The model is trained by fixing hyper-parameter settings of acoustic encoder layers $M_a = 16$, textual encoder layers $M_t = 5$, and three times of iteration in Sinkhorn attention. After the model is learned, CTC greedy search based decoding is used for recognition where only the components in acoustic modality is used. The results are showed in table 1. For comparison, the results of baseline system and several state-of-the-art systems which integrate BERT for linguistic knowledge transfer are also showed in Table 1. In this table, the 'Conformer-CTC' is the baseline system. 'Conformer-CTC/AED' denotes a hybrid CTC/AED ASR system as proposed in[3]. 'KT-RL-ATT' [9], and 'Wav2vec-BERT' [12] took pretrained acoustic model (from wav2vec2.0 [6]) and BERT for knowledge transfer. The two methods with 'Last-CMKT' or 'Hierarchical-CMKT' represents that our proposed CMKT was applied on the last hidden layer or on hierarchical of the acoustic encoders. From this table, we can see that our proposed CMKT yields competitive results. In particular, hierarchical CMKT achieved state of the art performance which suggested that linguistic knowledge transfer should be on both high and low-level of acoustic abstractions in order to improve ASR performance.

3.3. Ablation study

In this section, we figure out several important factors which affect the ASR performance in CMKT learning.

3.3.1. How many textual encoder layers are sufficient?

In textual encoder, several 'CM-encoder' layers are used to explore linguistic information with reference to features from acoustic encoder (refer to Fig. 2). With target linguistic representation from BERT as a supervision signal, the textual encoder could explore textual information from both textual and acoustic modalities. We did experiments with different number of CM-encoder layers, and showed results in table 2. From this table, we can see that it is necessary to increase the number of CM-encoder layers for the purpose of increasing textual encoder's capability to fully explore the information from textual and acoustic modalities.

Table 2. ASR performance with different number of CM-encoder layers, CER (%).

# CM-encoder layers	dev set	test set
$M_t = 1$	5.11	5.60
$M_t = 3$	3.69	4.05
$M_t = 5$	**3.64**	**3.94**
$M_t = 7$	3.66	3.99

Table 3. ASR performance with and without adapter connections in CMKT learning, CER (%).

Adapter connections	dev set	test set
Condition 1	5.25	5.77
Condition 2	4.20	4.54

3.3.2. Is the adapter necessary?

The adapter is a connection to pass acoustic information to text modality in CMKT learning, and transfer back the textual information conditioned on acoustic representations. Two experimental conditions are examined, i.e., Condition1: the adapter module is integrated in acoustic modality but no CMKT learning is performed (i.e., cut-off the connection to textual modality); Condition 2: the adapter is connected to textual modality with CMKT learning but it is not connected back to acoustic encoder (i.e., cut-off the FC3 link to acoustic modality in Fig. 2). The results are shown in table 3. From this table, we can see that the CMKT is the most important part in the proposed framework, and adapter with connections to both acoustic and textual modalities are also necessary.

4. CONCLUSION

In this study, we propose a novel hierarchical CMKT learning approach to enhance CTC-based ASR by harnessing linguistic representations encoded in a PLM model. CMKT learning involves transferring linguistic knowledge at both high and low levels of acoustic representations. In CMKT, we design Sinkhorn attention with just a few iterations to align cross-modal features. Using this alignment, the textual encoder can extract information from both textual and acoustic modalities to approximate the target linguistic representations encoded in BERT. By using an adapter that connects both acoustic and textual modalities, we efficiently transfer linguistic knowledge to the acoustic encoder. Our experiments confirm the effectiveness of the proposed CMKT learning framework.

The capacity of the proposed CMKT learning framework has not been fully explored. For example, questions remain regarding the integration of latent representations from the BERT model in CMKT learning and the adjustment of various hyperparameters in objective functions, especially concerning the Sinkhorn attention. In our future work, we will delve deeper into the potential of this learning framework through rigorous experimentation.

5. REFERENCES

[1] A. Graves, and N. Jaitly, "Towards end to-end speech recognition with recurrent neural networks," in *Proc. ICML*, pp. 17641772, 2014.

[2] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, DOI 10.1561/116.00000050, 2022.

[3] S. Watanabe, T. Hori, S. Kim, J. R. Hershey and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240-1253, 2017.

[4] Y. Higuchi, K. Karube, T. Ogawa, T. Kobayashi, "Hierarchical conditional end-to-end asr with ctc and multi-granular subword units," in *Proc. of ICASSP*, pp. 7797-7801, 2022.

[5] Y. Fujita, T. Komatsu, and Y. Kida, "Alternate Intermediate Conditioning with Syllable-Level and Character-Level Targets for Japanese ASR," in *Proc. of SLT*, pp. 76-83, 2022.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS*, 2020.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint* arXiv:1810.04805, 2018.

[8] M. Han, F. Chen, J. Shi, S. Xu, B. Xu, "Knowledge Transfer from Pre-trained Language Models to Cif-based Speech Recognizers via Hierarchical Distillation," *arXiv preprint* arXiv:2301.13003, 2023.

[9] K. Deng, S. Cao, Y. Zhang, L. Ma, G. Cheng, J. Xu, P. Zhang, "Improving CTC-Based Speech Recognition Via Knowledge Transferring from Pre-Trained Language Models," in *Proc. of ICASSP*, pp. 8517-8521, 2022.

[10] W. Cho, D. Kwak, J. Yoon, N. Kim, "Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation," in *Proc. of INTERSPEECH*, pp. 896-900, 2020.

[11] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen and S. Zhang, "Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring From BERT," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1897-1911, 2021.

[12] K. Lu and K. Chen, "A Context-aware Knowledge Transferring Strategy for CTC-based ASR," in *Proc. of SLT*, pp. 60-67, 2022.

[13] F. Yu, K. Chen, and K. Lu, "Non-autoregressive ASR Modeling using Pre-trained Language Models for Chinese Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1474-1482, 2022

[14] Y. Kubo, S. Karita, M. Bacchiani, "Knowledge Transfer from Large-Scale Pretrained Language Models to End-To-End Speech Recognizers," in *Proc. of ICASSP*, pp. 8512-8516, 2022.

[15] H. Futami, H. Inaguma, M. Mimura, S. Sakai, T. Kawahara, "Distilling the Knowledge of BERT for CTC-based ASR," CoRR abs/2209.02030, 2022.

[16] K. Choi, H. Park, "Distilling a Pretrained Language Model to a Multilingual ASR Model," in *Proc. of INTERSPEECH*, pp. 2203-2207, 2022.

[17] Y. Higuchi, T. Ogawa, T. Kobayashi, S. Watanabe, "BECTRA: Transducer-based End-to-End ASR with BERT-Enhanced Encoder," CoRR abs/2211.00792, 2022.

[18] C. Brodbeck, S. Bhattasali, A. Heredia, P. Resnik, J. Simon, E. Lau, "Parallel processing in speech perception with local and global representations of linguistic context," *Elife*, doi: 10.7554 /eLife .72056, 2022.

[19] X. Lu, P. Shen, Y. Tsao, H. Kawai, "Cross-Modal Alignment with Optimal Transport for CTC-Based ASR," IEEE-ASRU, Taipei, Taiwan, Dec.16-20, 2023.

[20] Y. Tay, D. Bahri, L. Yang, D. Metzler, D. Juan, "Sparse Sinkhorn Attention," in *Proc. of ICML*, pp. 9438-9447, 2020.

[21] M. Sander, P. Ablin, M. Blondel, G. Peyre, "Sinkformers: Transformers with Doubly Stochastic Attention," in *Proc. of AISTATS*, pp. 3515-3530, 2022.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NIPS*, pp. 5998-6008, 2017.

[23] C. Villani, Optimal transport: old and new, volume 338. Springer, 2009

[24] 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 *Proc. of COCOSDA*, pp. 1-5, 2017.

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

[26] D. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of ICLR*, 2015.