TEMPORAL ORDER PRESERVED OPTIMAL TRANSPORT-BASED CROSS-MODAL KNOWLEDGE TRANSFER LEARNING FOR 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

Transferring linguistic knowledge from a pretrained language model (PLM) to an acoustic model has been shown to greatly improve the performance of automatic speech recognition (ASR). However, due to the heterogeneous feature distributions in cross-modalities, designing an effective model for feature alignment and knowledge transfer between linguistic and acoustic sequences remains a challenging task. Optimal transport (OT), which efficiently measures probability distribution discrepancies, holds great potential for aligning and transferring knowledge between acoustic and linguistic modalities. Nonetheless, the original OT treats acoustic and linguistic feature sequences as two unordered sets in alignment and neglects temporal order information during OT coupling estimation. Consequently, a time-consuming pretraining stage is required to learn a good alignment between the acoustic and linguistic representations. In this paper, we propose a Temporal Order Preserved OT (TOT)-based Cross-modal Alignment and Knowledge Transfer (CAKT) (TOT-CAKT) for ASR. In the TOT-CAKT, local neighboring frames of acoustic sequences are smoothly mapped to neighboring regions of linguistic sequences, preserving their temporal order relationship in feature alignment and matching. With the TOT-CAKT model framework, we conduct Mandarin ASR experiments with a pretrained Chinese PLM for linguistic knowledge transfer. Our results demonstrate that the proposed TOT-CAKT significantly improves ASR performance compared to several state-of-the-art models employing linguistic knowledge transfer, and addresses the weaknesses of the original OT-based method in sequential feature alignment for ASR.

Index Terms— Optimal transport, Cross-modal knowledge transfer, automatic speech recognition

1. INTRODUCTION

The combination of a pretrained language model (PLM) with an end-to-end (E2E)-based acoustical model for automatic speech recognition (ASR) has made significant progress in recent years [1, 2, 3, 4, 5, 6]. The advantage of incorporating a PLM in ASR lies in the availability of unpaired large text corpora for training the PLM. Moreover, the linguistic knowledge encoded in the PLM can be utilized in ASR decoding. In most studies, the PLM is employed as an external language model (LM) for post-processing tasks such as beam search or rescoring in ASR [7, 8]. However, using an external LM for post-processing compromises the speed and sometimes parallel decoding capabilities of ASR. Addressing how to transfer linguistic knowledge to acoustic encoding during model training, and subsequently conducting speech recognition without relying on any external LM post-training, is an intriguing research topic.

In this study, our focus is on transferring linguistic knowledge from a PLM to a temporal connectionist temporal classification (CTC)-based ASR [9]. While there are several advanced end-to-end (E2E)-based ASR approaches that incorporate linguistic knowledge in acoustic model learning [10, 11], using a PLM, such as bidirectional encoder representation from transformers (BERT) [12]), facilitates linguistic knowledge transfer in ASR [13, 14, 15, 16], This knowledge transfer can also occur with a pretrained acoustic encoder, such as wav2vec2 [17], for both linguistic and acoustic knowledge transfer [18, 19, 20, 21, 22, 23, 24, 25]. However, due to the heterogeneous feature distributions in acoustic and linguistic spaces, it remains a challenging task to efficiently align feature representations between linguistic and acoustic modalities to facilitate knowledge transfer. In most studies, a cross-attention module is designed to integrate acoustic and text representations within a transformer decoder framework for combining acoustic and linguistic knowledge in ASR [26]. Yet, in the decoding stage, true text representations are unavailable, leading to the adoption of predicted text representations in decoding. This mismatch between training and testing phases weakens the benefits of linguistic information in ASR.

For efficient alignment and matching, an effective distance metric is needed to measure the difference between acoustic and linguistic feature representations. Considering this requirement, optimal transport (OT) emerges as a suitable tool for cross-modal alignment and linguistic knowledge transfer. OT, originally proposed for optimal allocating

Acknowledgment: The work is partially supported by JSPS KAKENHI No. 24K15004.

resources and later as a measure of discrepancies between probability distributions [27, 28], has found widespread applications in machine learning, particularly in domain adaptation [28]. In the field of speech, it has been employed for cross-domain spoken language recognition and speech enhancement [29, 30, 31], as well as in speech translation and understanding [32, 33, 34, 35]. The OT has been initially proposed for linguistic knowledge transfer learning for ASR in [36]. While OT is applicable for cross-domain alignment, its use in speech encounters a limitation. The original OT treats acoustic and linguistic feature sequences as two unordered sets in alignment and neglects temporal order information during OT coupling estimation. While acoustic and text speech exhibit a strong temporal order structure, requiring preservation of their temporal order relationship during alignment between an acoustic sequence and a linguistic sequence. Therefore, in [36], a well pretrained acoustic model was applied in order to efficiently explore matched acoustic features to those linguistic features during cross-modal learning. And the performance was strongly depended on the goodness of the pretrained acoustic model. In this paper, we propose a Temporal Order Preserved OT (TOT)-based Cross-modal Alignment and Knowledge Transfer (CAKT) model (TOT-CAKT) for CTC-based ASR. With the TOT-CAKT, the temporal order relationship is explicitly maintained during feature alignment and matching. It is hypothesized that through this alignment and matching, linguistic knowledge can be efficiently transferred to acoustic encoding, thereby enhancing ASR performance.

The rest of this paper is organized as follows: the proposed method is introduced in Section 2, where a cross-modal alignment module based on TOT and a neural adapter module for efficient linguistic feature transfer are designed. In Section 3, we conduct experiments to evaluate TOT-CAKT, comparing the results with several knowledge transfer learning algorithms for ASR, and provide a visualization of the learned transport coupling in OT. Finally, the conclusion is presented in Section 4.

2. PROPOSED METHOD

The model framework of the proposed TOT-CAKT method is illustrated in Fig. 1. This model framework is modified based on a conformer-CTC-based ASR model, incorporating two key modifications. First, an 'Adapter' module is added as shown in the gray blocks in Fig. 1. Second, an temporal order preserved OT-based cross-modal matching module is introduced in the right branch of Fig. 1. Both the acoustic features extracted from the conformer encoder and linguistic features derived from a PLM (with BERT utilized in this paper) are involved in the cross-modal matching process. Further details are provided in the following sections.

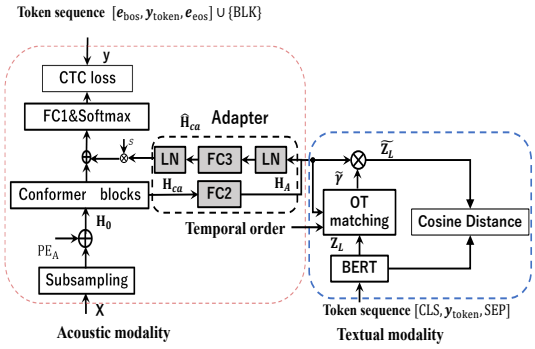

Fig. 1. The model framework of the proposed cross-modal alignment and knowledge transfer method based on TOT.

2.1. Acoustic and linguistic feature representations

The acoustic feature is extracted from the acoustic encoder where a conformer-based encoder [37]) is adopted. The process in the 'Subsampling' module involves a two-layer convolution process with a downsampling operation (a downsampling rate of 4 was used in this paper). By incorporating a positional encoding from $\mathrm{PE_A}$, the initial input to conformer blocks is obtained as $\mathbf{H}_0$. The output of the conformer encoder is represented as an acoustic representation $\mathbf{H}_{\mathrm{ca}}$.

$$\begin{aligned} \mathbf{H}_0 &= \mathrm{Subsampling}\left(\mathbf{X}\right) + \mathrm{PE_A} \\ \mathbf{H}_{\mathrm{ca}} &= \mathrm{Conformer}\left(\mathbf{H}_0\right) \in R^{l_a \times d_a}, \end{aligned} \tag{1}$$

where l_a and d_a are temporal length and dimension of the acoustic feature vectors, respectively. Before engaging in cross-modal feature alignment, a linear projection termed 'FC2' in the 'Adapter' module is utilized to perform a feature dimension matching transform:

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

In this equation, d_t corresponds to linguistic feature dimension. In the right branch of Fig. 1, the context-dependent linguistic feature representation is explored from a pretrained BERT model. The process is formulated as:

$$\begin{aligned} \mathbf{y}_{\mathrm{token}} &= \mathrm{BERTTokenizer}\left(\mathbf{y}\right) \\ \mathbf{Z}_0 &= [\mathrm{CLS}, \mathbf{y}_{\mathrm{token}}, \mathrm{SEP}] \\ \mathbf{Z}_i &= \mathrm{BERT}_i\left(\mathbf{Z}_{i-1}\right), \end{aligned} \tag{3}$$

where 'BERT_i' is the i-th transformer encoder layer of BERT model, i takes values from 1 to L, with L representing the total number of BERT encoder layers. 'BERTTokenizer' is a process to convert standard text to word piece based tokens [12]. Token symbols 'CLS' and 'SEP' represent the start and end of an input sequence. $\mathbf{Z}_L \in \mathbb{R}^{l_t \times d_t}$ is the final text representation which encodes context dependent linguistic information, l_t denotes the sequence length, and d_t represents feature dimension of text encoding representation.

2.2. Sinkhorn algorithm for cross-modal alignment

The original OT was formulated to transform from one probability distribution to another with minimum transport cost [27]. In this study, we applied OT for feature alignment on two sets. Given acoustic and linguistic feature sequences $\mathbf{H}_A$ and $\mathbf{Z}_L$ respectively, as:

$$\begin{array}{l}\mathbf{H}_A = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_i, ..., \mathbf{h}_{l_a}] \\ \mathbf{Z}_L = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_j, ..., \mathbf{z}_{l_t}],\end{array} \tag{4}$$

where l_a and l_t are lengths of the two sequences. Suppose the two sequences in Eq. (4) are sampled from two probability distributions with weight vectors $\mathbf{a} = [a_1, a_2, ..., a_i, ..., a_{l_a}]$ and $\mathbf{b} = [b_1, b_2, ..., b_j, ..., b_{l_t}]$. ($a_i = 1/l_a, b_j = 1/l_t$ as uniform distributions if no prior information is available). The OT distance between the two sequences is defined as:

$$L_{\mathrm{OT}} \triangleq \min_{\gamma \in \prod(\mathbf{H}_A, \mathbf{Z}_L)} \langle \gamma, \mathbf{C} \rangle_{F.}, \tag{5}$$

where γ is a transport coupling set defined as:

$$\prod(\mathbf{H}_A, \mathbf{Z}_L) \triangleq \left\{ \gamma \in R_+^{l_a \times l_t} \left| \gamma \mathbf{1}_{l_t} = \mathbf{a}, \gamma^T \mathbf{1}_{l_a} = \mathbf{b} \right. \right\} \tag{6}$$

In Eq. (6), $\mathbf{1}_{l_a}$ and $\mathbf{1}_{l_t}$ are vectors of ones with dimensions l_a and l_t, respectively, γ is a non-negative coupling matrix with dimensions $l_a \times l_t$. In Eq. (5), $\langle , \rangle_{F.}$ is the Frobenius inner product operator on two matrices, $\mathbf{C}$ is a distance matrix (or ground metric) with element $c_{i,j}$ defined as pair-wised cosine distance:

$$c_{i,j} = \mathbf{C}(\mathbf{h}_i, \mathbf{z}_j) \triangleq 1 - \cos(\mathbf{h}_i, \mathbf{z}_j) \tag{7}$$

A fast estimation of OT has been introduced through the celebrated entropy-regularized OT (EOT) [38] where the EOT loss is defined as:

$$L_{\mathrm{EOT}}(\mathbf{H}_A, \mathbf{Z}_L) \triangleq \min_{\gamma \in \prod(\mathbf{H}_A, \mathbf{Z}_L)} \langle \gamma, \mathbf{C} \rangle_{F.} - \alpha_1 H(\gamma), \tag{8}$$

where α_1 is a regularization coefficient, and $H(\gamma)$ is entropy of coupling matrix defined as:

$$H(\gamma) = -\sum_{i,j} \gamma_{i,j} \log \gamma_{i,j}. \tag{9}$$

The solution of Eq. (8) can be implemented with Sinkhorn algorithm as [27]:

$$\gamma_{\alpha_1} = diag(\mathbf{u}_1) * \mathbf{G} * diag(\mathbf{u}_2) \tag{10}$$

where $\mathbf{G} = \exp\left(-\frac{\mathbf{C}}{\alpha_1}\right)$, $\mathbf{u}_1$ and $\mathbf{u}_2$ are two scaling (or re-normalization) vectors.

2.3. Temporal order preserved OT

In the original estimation of OT in Eq. (8), the two sequences in Eq. (4) are treated as two sets without considering their temporal order relationship. In speech, temporal order information is crucial in OT coupling during cross-modal alignment, meaning that neighboring frames in an acoustic sequence should be progressively coupled with the neighboring tokens in a linguistic sequence. Therefore, as showed in Fig. 1, the temporal order information is input to the OT matching block. For the sake of clarity, the two sequences in Eq. (4) can be further represented with temporal order information as:

$$\begin{array}{l}\mathbf{H}_A = [(\mathbf{h}_1, 1), (\mathbf{h}_2, 2), ..., (\mathbf{h}_i, i), ..., (\mathbf{h}_{l_a}, l_a)] \\ \mathbf{Z}_L = [(\mathbf{z}_1, 1), (\mathbf{z}_2, 2), ..., (\mathbf{z}_j, j), ..., (\mathbf{z}_{l_t}, l_t)]\end{array} \tag{11}$$

During the alignment of the two sequences for knowledge transfer, it is crucial to consider that elements with significant cross temporal distances might not be likely to be coupled. In other words, the coupling pairs with high probabilities between the two sequences should be distributed along the diagonal line of the temporal coherence positions. Based on this consideration, the temporal coupling prior could be defined as a two dimensional Gaussian distribution [39]. The fundamental concept is that the coupled pairs should not deviate significantly from the diagonal line of temporal coherence positions between the two sequences, which can be defined as:

$$p_{i,j} \triangleq \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{d_{i,j}^2}{2\sigma^2}\right), \tag{12}$$

where σ is a variation variable controlling the impact of the cross-temporal distance $d_{i,j}$ as defined in Eq. (13).

$$d_{i,j} = \frac{|i/l_a - j/l_t|}{\sqrt{1/l_a^2 + 1/l_t^2}} \tag{13}$$

In Eq. (13), the cross-temporal distance is defined on the normalized sequence lengths in acoustic and linguistic spaces. In this definition, it is evident that the farther the distance between a paired position and the temporal diagonal line, the lower possibility of their correspondence in transport coupling. By incorporating this temporal coherence prior as regularization, the new OT is defined as:

$$L_{\mathrm{TOT}}(\mathbf{H}_A, \mathbf{Z}_L) \triangleq \min_{\gamma \in \prod(\mathbf{H}_A, \mathbf{Z}_L)} \langle \gamma, \mathbf{C} \rangle_{F.} - \alpha_1 H(\gamma) + \alpha_2 KL(\gamma || P), \tag{14}$$

where α_1 and α_2 are two trade off parameters. In Eq. (14), $KL(\gamma||P)$ is the Kullback-Leibler (KL) divergence between the transport coupling matrix γ and temporal prior correspondence matrix P with elements defined in Eq. (12). Building upon the definitions of KL-divergence and entropy, Eq. (14) can be further expressed to:

$$L_{\mathrm{TOT}}(\mathbf{H}_A, \mathbf{Z}_L) \triangleq \min_{\gamma \in \prod(\mathbf{H}_A, \mathbf{Z}_L)} \left\langle \gamma, \tilde{\mathbf{C}} \right\rangle_{F.} - \tilde{\alpha} H(\gamma), \tag{15}$$

where $\tilde{\alpha} = \alpha_1 + \alpha_2$, and combined ground cost matrix as

$$\tilde{\mathbf{C}} = \mathbf{C} - \alpha_2 \log P \tag{16}$$

Following the procedures outlined in [38], the solution of Eq. (15) is obtained using the Sinkhorn algorithm as:

$$\gamma_{\tilde{\alpha}} = diag\left(\mathbf{u}_1\right) * \tilde{\mathbf{G}} * diag\left(\mathbf{u}_2\right), \tag{17}$$

where $\tilde{\mathbf{G}} = \exp\left(-\frac{\tilde{\mathbf{C}}}{\tilde{\alpha}}\right)$. Substituting variables in Eq. (16) to $\tilde{\mathbf{G}}$, we can obtain:

$$\tilde{\mathbf{G}} = P^{\frac{\alpha_2}{\alpha_1+\alpha_2}} \exp(-\frac{\mathbf{C}}{\alpha_1+\alpha_2}) \tag{18}$$

From this equation, we can see that the transport coupling between the two sequences is further constrained by their temporal order correspondence. TOT involves several hyper-parameters that can be challenging to control. For the sake of simplification, we consolidate their effects into a reduced number of hyper-parameters. For example, considering Eq. (16), the impact of variation σ in Eq. (12) and α_2 in Eq. (14) can be combined into a single control parameter β, defined as:

$$\tilde{\mathbf{C}} = \mathbf{C} + \beta d_{i,j}^2 \tag{19}$$

And the Sinkhorn algorithm is applied on the cost function matrix $\tilde{\mathbf{C}}$ for OT in real implementations.

2.4. Loss function

The proposed TOT-CAKT involves two loss functions: the cross-modal alignment and matching loss (in the right branch of Fig. 1) and the CTC loss (in the left branch of Fig. 1). In cross-modal alignment, the acoustic feature can be projected onto the linguistic space using OT as:

$$\begin{aligned} \tilde{\mathbf{Z}}_L &\triangleq \mathrm{OT}\left(\mathbf{H}_A \rightarrow \mathbf{Z}_L\right) \\ &= \gamma^* \times \mathbf{H}_A \in R^{l_t \times d_t}, \end{aligned} \tag{20}$$

where γ^* is the optimal transport coupling based on OT. Subsequently, the alignment loss is defined as:

$$L_{\text{align}} = \sum_{j=2}^{l_t-1} 1 - \cos\left(\tilde{\mathbf{z}}_L^j, \mathbf{z}_L^j\right), \tag{21}$$

where $\tilde{\mathbf{z}}_L^j$ and $\mathbf{z}_L^j$ are row vectors of feature matrices $\tilde{\mathbf{Z}}_L$ and $\mathbf{Z}_L$ (matching on temporal dimensions), respectively. In Eq. (21), the usage of indices from 2 to $l_t - 1$ is for handling special symbols 'CLS' and 'SEP'. For efficient linguistic knowledge transfer to acoustic encoding, the following transforms are designed as indicated in Fig. 1:

$$\begin{aligned} &\hat{\mathbf{H}}_{\text{ca}} = \mathrm{FC}_3\left(\mathrm{LN}(\mathbf{H}_A)\right) \in R^{l_a \times d_a} \\ &\mathbf{H}^{a,t} = \mathbf{H}_{\text{ca}} + \mathrm{s} \cdot \mathrm{LN}(\hat{\mathbf{H}}_{\text{ca}}), \end{aligned} \tag{22}$$

where s is a scaling parameter to adjust the importance of transferring linguistic projected feature, 'LN' is layer normalization operator. Based on this new representation $\mathbf{H}^{a,t}$ which is intended to encode both acoustic and linguistic information, the final probability prediction for ASR is formulated as:

$$\tilde{\mathbf{P}} = \mathrm{Softmax}\left(\mathrm{FC1}\left(\mathbf{H}^{a,t}\right)\right), \tag{23}$$

where 'FC1' is a linear full-connected transform. The total loss in model learning is defined as:

$$L \triangleq \lambda . L_{\text{CTC}}(\tilde{\mathbf{P}}, \mathbf{y}_{\text{token}}) + (1-\lambda) . w . (L_{\text{align}} + L_{\text{TOT}}), \tag{24}$$

where $L_{\text{CTC}}(\tilde{\mathbf{P}}, \mathbf{y}_{\text{token}})$ is CTC loss, L_{align} and L_{TOT} are cross-modality alignment loss and TOT loss, respectively. After the model is trained, only the left branch of Fig. 1 is retained for ASR inference.

3. EXPERIMENTS

ASR experiments were conducted on the open-source Mandarin speech corpus AISHELL-1 [40] to evaluate the proposed algorithm. The data corpus comprises three datasets: 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). Data augmentation as used in [40] was applied. Given the tonal nature of the Mandarin language in the ASR task, in addition to using 80-dimensional log Mel-filter bank features, three extra acoustic features related to fundamental frequency, i.e., F0, delta F0 and delta delta F0, were utilized as raw input features. These features were extracted with a 25ms window size and a 10ms shift.

3.1. Model architecture

In Fig. 1, the 'Subsampling' module consists of two CNN blocks with 256 channels, kernel size 3, stride 2, and ReLU activation function in each. The acoustic encoder is formed by stacking 16 conformer blocks [37], with each having a kernel size of 15, attention dimension $d_a = 256$, 4 attention heads, and a 2048-dimensional FFN layer. The 'bert-base-chinese' from huggingface is used as the pretrained PLM for linguistic knowledge transfer [41]. In this Chinese BERT model, 12 transformer encoders are applied, the token (or vocabulary) size is 21128, and the dimension of linguistic feature representation is $d_t = 768$.

3.2. Hyper-parameters in model learning

Several hyper-parameters are associated with the proposed model, and these parameters may have a joint (or correlated) effect in efficient linguistic knowledge transfer learning. In our preliminary experiments, for easy implementation, they were fixed as $\beta = 0.5$ in Eq. (19), alignment trade off parameter $\lambda = 0.3$ and scale parameter $w = 1.0$ in Eq. (24).

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

Methods	dev set	test set
Conformer+CTC (Baseline)	5.53	6.05
Conformer+CTC/AED ([5, 43])	4.61	5.06
NAR-BERT-ASR ([14])	4.90	5.50
KT-RL-ATT ([22])	4.38	4.73
Wav2vec-BERT ([21])	4.10	4.39
Without pretraining condition		
OT-BERT(w/o) ([36])	4.21	4.52
TOT-CAKT(w/o),$\tilde{\alpha} = 0.01$,$s = 1.0$	4.03	4.30
TOT-CAKT(w/o), $\tilde{\alpha} = 0.1$,$s = 1.0$	4.16	4.48
TOT-CAKT(w/o),$\tilde{\alpha} = 0.5$,$s = 1.0$	3.99	4.36
TOT-CAKT(w/o),$\tilde{\alpha} = 0.5$,$s = 0.5$	4.06	4.40
TOT-CAKT(w/o),$\tilde{\alpha} = 0.5$,$s = 0.1$	**3.93**	4.35
TOT-CAKT(w/o),$\tilde{\alpha} = 0.01$,$s = 0.1$	4.02	**4.29**
With pretraining condition		
OT-BERT(w/)([36])	3.96	4.27
TOT-CAKT(w/)	**3.88**	**4.21**

$\tilde{\alpha}$ in Eq. (15) and s in Eq. (22) were varied in experiments. For optimization, Adam optimizer [42] is used with a learning rate (initially set to 0.001) schedule with 20,000 warm-up steps. The model with cross-modal transfer was trained for 130 epochs, and the final model used for evaluation was obtained by averaging models from the last 10 epochs (the original conformer-CTC model without the adapter module was pretrained or trained from scratch in different experimental settings when jointly combined with cross-modal learning). The performance was evaluated based on character error rate (CER).

3.3. Results

In inference stage, only the left branch (blocks in dashed red box in Fig. 1) is utilized, maintaining the decoding speed similar to that of the CTC-based decoding. In our experiments, only CTC greedy search-based decoding was employed, and the results are presented in table 1. The results of the baseline system and several state-of-the-art systems that integrate BERT for linguistic knowledge transfer are also provided for comparison. In this table, 'Conformer+CTC' is the baseline system, trained without linguistic knowledge transfer. 'Conformer+CTC/AED' denotes a hybrid CTC/AED ASR system [3, 4, 5] which used a transformer decoder with attention to text representation during model training. 'NAR-BERT-ASR', 'KT-RL-ATT', and 'Wav2vec-BERT' are all based on integrating acoustic and linguistic features from BERT for ASR [22, 23, 14, 21], and even used a pretrained acoustic model (from wav2vec2.0 [17]) and PLM for knowledge transfer. In the OT based cross-modal learning, two experimental conditions were examined. One is that models with cross-modal learning were trained from scratch, i.e., without pretraining condition. The other is that the models were initialized with a pretrained acoustic model, then were further trained with cross-modal learning, i.e., with pretraining condition. Correspondingly, in table 1, 'OT-BERT(w/o)' and 'TOT-CAKT(w/o)' denote results of cross-modal linguistic knowledge transfer learning based on OT in [36] and the proposed TOT-CAKT method for without pretraining condition, respectively, 'OT-BERT(w/)' and 'TOT-CAKT(w/)' are results with pretraining condition. From this table, we can observe that linguistic knowledge significantly enhances the ASR performance (not in a conventional way like LM rescoring). From the results in 'OT-BERT(w/o)' and 'TOT-CAKT(w/o)', both the methods in [36] and the proposed cross-modal learning could efficiently transfer linguistic knowledge in acoustic encoding, yielding promising results. Moreover, when comparing results in 'OT-BERT(w/o)' and 'TOT-CAKT(w/o)', a significant performance improvement was observed when all models with cross-modal learning were trained from scratch (without pretraining condition). Furthermore, the performance of our proposed TOT-CAKT, even without pretraining could reach a level comparable to the original OT-based method, which requires a time-consuming pretraining stage (as in 'OT-BERT(w/)). Finally, we further examined that when pretraining was utilized before cross-modal learning, the improvement of our proposed method (as shown in 'TOT-CAKT(w/)' in table 1) compared to the 'OT-BERT(w/)' in [36] was reduced. This suggests that the pretraining stage could implicitly provide temporal order information for cross-modal feature alignment and knowledge transfer. In comparison, our proposed method explicitly incorporates temporal order information in the mathematical modeling with flexible parameters for control in experiments. Based on our formulation, our future work will focus on finding optimal temporal order parameter settings.

3.4. Visualization of transport coupling

In the proposed TOT-CAKT, the coupled pairs between the acoustic and linguistic feature sequences are explicitly designed to correspond to their temporal coherence, i.e., acoustic segments should match well with their linguistic tokens sequentially. Two examples of the coupling matrices are shown in Fig. 2. In Fig. 2-a, the coupling matrix is learned based on OT without temporal order constraint, and Fig. 2-b is the coupling matrix learned with temporal order constraint. From this figure, it is evident that clear temporal correspondences exist between the acoustic feature sequence and linguistic token sequence in both transport couplings. Moreover, several positions with incorrect couplings in Fig. 2-a were corrected in Fig. 2-b by our proposed method, which explicitly adds a temporal order constraint.

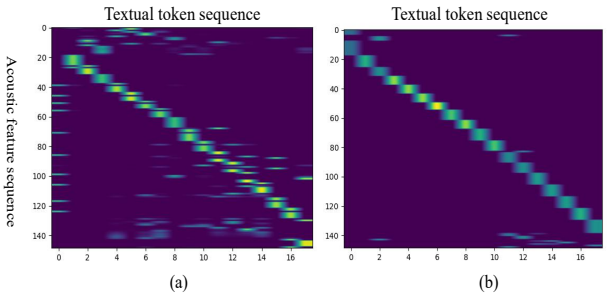

Fig. 2. Transport coupling matrix without temporal order constraint (a); with temporal order constraint (b). (Both of these cross-modal learning methods started from scratch without acoustic model pretraining.)

4. CONCLUSION

Acoustic and linguistic features belonging to different modalities require feature alignment as a crucial step in transferring linguistic knowledge from a PLM to acoustic encoding. In this paper, we proposed a novel TOT-CAKT. In the TOT-CAKT, a transfer coupling or mapping preserving temporal order information between acoustic sequence and linguistic sequence was first estimated. Subsequently, acoustic features were mapped to the linguistic space based on the transport coupling, allowing direct comparison of the mapped features to match the information encoded in the PLM. Our ASR experimental results confirmed the effectiveness of the proposed TOT-CAKT. Additionally, based on the visualization of transport coupling, we verified that TOT could eliminate unreasonable matches between the acoustic and linguistic sequences.

In the proposed TOT-CAKT, several hyper-parameters involved in model learning are challenging to control. Additionally, some of these hyper-parameters are sensitive in implementation and may lead to stability problems in the Sinkhorn algorithm. In the current paper, the combined effects of those hyper-parameters have not been clearly explored, and a clear understanding of their combined rules has not been established, potentially hindering the identification of optimal solutions for improving ASR performance. Figuring out a set of optimal hyper-parameters in the proposed TOT-CAKT remains as our future work.

5. REFERENCES

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

[2] W. Chan, N. Jaitly, Q. Le and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of ICASSP*, 2016, pp. 4960-4964.

[3] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. of ICASSP*, 2017, pp. 48354839.

[4] T. Hori, S. Watanabe, and J. R. Hershey, "Joint ctc/attention decoding for end-to-end speech recognition," in *Proc. of ACL*, 2017, vol. 1, pp. 518529.

[5] 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.

[6] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint*, arXiv:1211.3711, 2012.

[7] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using BERT for speech recognition," in *Proc. of ACML*, 2019, pp. 1081-1093.

[8] J. Salazar, D. Liang, T. Nguyen, K. Kirchhoff, "Masked Language Model Scoring," in *Proc. of ACL*, 2020, pp. 2699-2712.

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

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

[11] Y. Fujita, T. Komatsu, and Y. Kida, "Multi-sequence intermediate conditioning for ctc-based asr," *arXiv preprint*, arXiv:2204.00175, 2022.

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

[13] 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.

[14] 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

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

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

[17] 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.

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

[19] Y. Higuchi, T. Ogawa, T. Kobayashi, S. Watanabe, "BECTRA: Transducer-Based End-To-End ASR with Bert-Enhanced Encoder," in *Proc. of ICASSP*, 2023, pp. 1-5.

[20] 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.

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

[22] 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*, 2022, pp. 8517-8521.

[23] K. Deng, Z. Yang, S. Watanabe, Y. Higuchi, G. Cheng, P. Zhang, "Improving Non-Autoregressive End-to-End Speech Recognition with Pre-Trained Acoustic and Language Models," in *Proc. of ICASSP*, 2022, pp. 8522-8526.

[24] K. Deng, Z, P. Woodland, "FastInject: Injecting Unpaired Text Data into CTC-Based ASR Training," in *Proc. of ICASSP*, 2024, pp. 11836-11840.

[25] M. Han, L. Dong, Z. Liang, M. Cai, S. Zhou, Z. Ma, B. Xu, "Improving End-to-End Contextual Speech Recognition with Fine-Grained Contextual Knowledge Selection," in *Proc. of ICASSP*, 2022, pp. 8532-8536.

[26] 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*, 2017, pp. 5998-6008.

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

[28] N. Courty, R. Flamary, A. Habrard, A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Proc. of NIPS*, 2017, pp. 3733-3742.

[29] X. Lu, P. Shen, Y. Tsao, H. Kawai, "Unsupervised Neural Adaptation Model Based on Optimal Transport for Spoken Language Identification," in *Proc. of ICASSP*, 2021, pp. 7213-7217.

[30] H. Lin, H. Tseng, X. Lu, Y. Tsao, "Unsupervised Noise Adaptive Speech Enhancement by Discriminator-Constrained Optimal Transport," in *Proc. of NeurIPS*, 2021, pp. 19935-19946.

[31] H. Tseng, H. Lin, H. Hsuan, and Y. Tsao, "Interpretations of Domain Adaptations via Layer Variational Analysis," *arXiv preprint*, CoRR abs/2302.01798, 2023.

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

[33] W. Wang, S. Ren, Y. Qian, S. Liu, Y. Shi, Y. Qian, M. Zeng, "Optimizing Alignment of Speech and Language Latent Spaces for End-To-End Speech Recognition and Understanding," in *Proc. of ICASSP*, 2021, pp. 7802-7806.

[34] Y. Zhou, Q. Fang, Y. Feng, "CMOT: Cross-modal Mixup via Optimal Transport for Speech Translation," *arXiv preprint*, arXiv:2305.14635, 2023.

[35] P. Le, H. Gong, C. Wang, J. Pino, B. Lecouteux, D. Schwab, "Pre-training for Speech Translation: CTC Meets Optimal Transport," *arXiv preprint*, CoRR abs/2301.11716, 2023.

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

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

[38] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. of NIPS*, 2013, vol. 26.

[39] B. Su, G. Hua, "Order-Preserving Wasserstein Distance for Sequence Matching," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[40] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "AIShell-1: An open-source mandarin speech corpus and a speech recognition baseline, in *Proc. of COCOSDA*, 2017, pp. 1-5.

[41] https://huggingface.co/

[42] Diederik P. Kingma, Jimmy Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of ICLR*, 2015.

[43] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, J. Niu, "WeNet 2.0: More Productive End-to-End Speech Recognition Toolkit," in *Proc. of INTERSPEECH*, 2022, pp. 1661-1665.