IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2020, accepted January 27, 2020, date of publication January 31, 2020, date of current version February 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2970758

Towards Understanding Attention-Based Speech Recognition Models

CHU-XIONG QIN AND DAN QU

Department of Information and Systems Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China

Corresponding author: Dan Qu (qudan2019@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61673395 and Grant 61403415, and in part by the Natural Science Foundation of Henan Province under Grant 162300410331.

ABSTRACT Although the attention-based speech recognition has achieved promising performances, the specific explanation of the intermediate representations remains a black box theory. In this paper, we use the method to visually show and explain continuous encoder outputs. We propose a human-intervened force alignment method to obtain labels for t-distributed stochastic neighbor embedding (t-SNE), and use them to better understand the attention mechanism and the recurrent representations. In addition, we combine t-SNE and canonical correlation analysis (CCA) to analyze the training dynamics of phones in the attention-based model. Experiments are carried on TIMIT and WSJ respectively. The aligned embeddings of the encoder outputs could form sequence manifolds of the ground truth labels. Figures of t-SNE embeddings visually show what representations the encoder shaped into and how the attention mechanism works for the speech recognition. The comparisons between different models, different layers, and different lengths of the utterance show that manifolds are clearer in the shape when outputs are from the deeper layer of the encoder, the shorter utterance, and models with better performances. We also observe that the same symbols from different utterances tend to gather at similar positions, which proves the consistency of our method. Further comparisons are taken between different epochs of the model using t-SNE and CCA. The results show that both the plosive and the nasal/flap phones converge quickly, while the long vowel phone converge slowly.

INDEX TERMS Attention-based model, t-distributed stochastic neighbor embedding, canonical correlation analysis.

I. INTRODUCTION

The traditional techniques separate a speech recognition system into a variety of modules. The system which is designed by these techniques is grounded with many assumptions, and some of the modules require expert knowledge. For example, the acoustic model is trained in a frame-wise manner which is based on the Markov assumption [1], and the decoding stage needs a man-made dictionary to obtain hypothesizes [2].

To eliminate all potential unreasonable artificial designs in the system, there raised a variety of methods that model speech signals in an end-to-end way. The main task of an end-to-end model is to create mappings between two different sequences (input features and sequences of symbols) with different lengths. The connectionist temporal classification (CTC) model [3], [4] and the attention-based model [5]–[7] are two typical successful explorations.

The CTC model inserts extra blank symbols to make the length of the outputs consistent with the input sequences. Although it is optimized at a sequence-level, it is still a variant of the Markov model which is based on an independent assumption [8].

The attention-based model is another important extension of the end-to-end approach. It is always composed of an encoder, an attention layer, and a decoder. The attention mechanism is considered as mappings from the outputs of the encoder to the inputs of the decoder states [9]–[11]. Since the number of input acoustic features is much larger than the number of symbols in the label sequence, the attention weights are reflected as many-to-one connections in the speech recognition.

Although there are other variants of end-to-end models such as the transformer [12] and the neural

The associate editor coordinating the review of this manuscript and approving it for publication was Stavros Ntalampiras.

transducer [13], [14], the attention layer is a basic structure for many end-to-end research. It is more interpretable than other methods, and it achieves rather promising results [5], [15], [16].

Nevertheless, the attention-based model and its variants are based on the deep learning theory, which is still a black box in many applications. In image processing, people have successfully uncovered a lot of interpretations on convolutional neural network (CNN) [17]–[19]. However, understanding intermediate outputs in speech recognition is challenging. Speech signals only have short-term stability and are always framed before modeling. Besides, there are too many variables embedded in speech signals, making it almost impossible to recover the original speech signals from transformed representations. As for the attention-based model, we would like to explore the outputs from intermediate layers of recurrent neural network (RNN) and why attention mechanism works for speech recognition.

II. RELATED WORKS

There have been many explorations toward interpreting neural network outputs in the speech processing area. Bai et al. [20] proposed to use linear discriminant analysis (LDA) and t-distributed stochastic neighbor embedding (t-SNE) to analyze 9-dimensional bottleneck features (BNFs). Karita et al. [21] use t-SNE to visually show how features are mixed or split with inter-domain loss. Kim et al. [22] use t-SNE on high-level features to show the distribution of emotional categories. Tang et al. [23] show temporal traces of recurrent units with t-SNE at different layers. Google Brain uses singular vector canonical correlation analysis (SVCCA) [24] and projection weighted CCA [25] to compare representational similarity between two different CNNs or deep neural networks (DNNs), for better understanding of the deep learning dynamics. Zhou *et al.* [26] proposes to use finite state automaton (FSA) to learn intermediate output structures of RNN.

Most of the previous research interpret on neural networks within traditional techniques, with only a few researches aiming for the end-to-end structure.

To better understand the intermediate representations and the training dynamics of the attention-based model. First, we apply t-SNE to the encoder outputs. Then, in order to visualize those embeddings, we propose a human-intervened force alignment method to obtain labels in the frame-level. Finally, we try to understand the training dynamics using CCA upon t-SNE embeddings. The analyze is done in a phone-level.

We experiment on TIMIT and WSJ. The drawings of t-SNE embeddings all show that the encoder of the attention-based model clusters similar data points by the class of symbols and form a manifold graph of sequential symbols. We further experiment on comparing training dynamics using a combined method of t-SNE and CCA. It shows that phones with long tones are learned quickly while phones with short tones converge slowly.

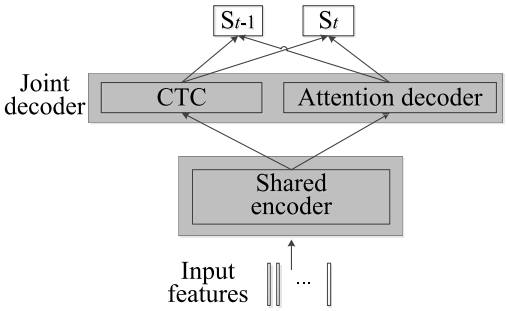


FIGURE 1. The structure of the joint CTC-attention model.

The structure is organized as follows: Section III introduces our end-to-end models. Section IV describes the algorithms of t-SNE and our force alignment method for the attention-based model. Section V introduces how we use CCA to analyze the training dynamics of the phones. Section VI presents our experiments and analyzes the experimental results. Finally, Section VII concludes the paper and points out the potential future work.

III. MODEL DESCRIPTIONS

The target model we analyze in this paper is a joint CTC-attention model. We study this hybrid model instead of a pure attention-based model because it achieves better results and shows great potential in the end-to-end speech recognition [27]–[29]. In this section, we describe the model briefly.

The joint model combines the CTC and the attention-based model. It is composed of a shared encoder and a joint decoder, and the structure is shown in Fig. 1.

The input features are normally acoustic features such as MFCCs and filter banks. In our previous research, we demonstrate that the performance gets better using high-level features [29], [30].

In this paper, the components of the encoder are made of bi-directional long short-term memory (BLSTM) layers and convolutional layers. This follows configurations in typical researches like [31], [32], so that our research could be representative. The encoder transforms the input matrix X into representations H by:

$$\boldsymbol{H} = Encoder\,(\boldsymbol{X}) \tag{1}$$

For the joint CTC-attention model, the loss function $\mathcal{L}$ is a linear combination of two parts:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CTC} + (1 - \lambda_1)\,\mathcal{L}_{atermon}, \quad \lambda \in [0, 1] \tag{2}$$

where λ_1 is the linear weight of CTC loss. $\mathcal{L}_{CTC}$ and $\mathcal{L}_{attention}$ are the losses for the CTC and the attention-based model respectively.

Finally, the joint decoding also uses CTC to help provide the most probable phone sequence $\boldsymbol{S}$:

$$\boldsymbol{S} = \arg\max \{\lambda_2 \alpha_{CTC} + (1 - \lambda_2)\,\alpha_{att}\} \tag{3}$$

where λ_2 is the linear weight of CTC. $\alpha_{\boldsymbol{CTC}}$ and $\alpha_{\boldsymbol{att}}$ are the hypothesis output from the CTC and the attention-based model.

IV. VISUALIZING ENCODER OUTPUTS USING T-SNE AND FORCE ALIGNMENT

In an attention-based model, the attention layer is the most interpretable part, as it tries to align between different representations in sequential order. The decoder part is usually a shallow network composed of few LSTM layers, and it predicts phone/character sequences. For the encoder, it maps the input features $\boldsymbol{X}$ into hidden representations $\boldsymbol{H}$, which we still have no clear analysis of what they really represent. In this chapter, our proposed method uses t-SNE and force alignment to help understand the encoder of the attention-based model.

The encoder is usually composed of a few stacked BLSTM layers. On some occasions, there are some convolutional layers at the bottom of the network, but high layers are always BLSTM. Even if the network is applied with resolution reduction, it still outputs high-dimensional vectors. In order to visualize those representations, we use t-SNE to narrow the output layer to 2 nodes to have a direct view of an encoder space.

Upon SNE, t-SNE could alleviate both the crowding problem and the optimization problems. Denote by $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2 \ldots, \boldsymbol{h}_T\}$ where T is the number of data points. Each of the vectors is K-dimensional output from the encoder.

Denote $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2 \ldots, \boldsymbol{y}_T\}$ as the set of low-dimensional t-SNE embeddings (with a size of $T \times 2$), it is obtained by:

$$\boldsymbol{Y} = \mathrm{TSNE}(\boldsymbol{H}) \tag{4}$$

Here $\mathrm{TSNE}(\cdot)$ represents for the t-SNE transformation operation. The details of the t-SNE algorithm could be found in [33], and we do not describe it here.

To clear visualize all the t-SNE embeddings and find patterns for them, the labels of all data points are required. However, there is no explicit alignment for each frame in the attention-based model since the loss is optimized in an utterance-level.

In this section, we propose a modified force alignment method for the attention-based model. The alignments come from the attention weights. Given an attention weights matrix $\boldsymbol{W}$ which is organized by T rows and N columns in an attention-based model, T and N respectively represent for the number of input features (data points) and the number of symbols in one utterance. First, we apply a greedy search to the weights matrix to obtain raw alignments. Denote by $\boldsymbol{L} = (L_1, L_2 \ldots L_N)$ where $\boldsymbol{L}$ is the raw alignments and L_i is the ith symbol in $\boldsymbol{L}$. Then L_i is calculated as:

$$L_i = \arg\max_j (W(i, j)) \tag{5}$$

where $W(i, j)$ is the element in the ith row and the jth column of $\boldsymbol{W}$.

However, there are mistakes that are mainly located at the beginning and the ending of a label sequence due to padding. We take an example from TIMIT, where the labels for each utterance are started and ended with "sil". The ground truth label of "sil eh n w ah dx ay z dh ey w er sil" will become "sos sil eh n w ah dx ay z dh ey w er sil" after adding a symbol of "sos" at the beginning when it is predicted from an attention-based model. The beginning of the raw alignments looks like "sos sos sos sos sil sil sos sos sil sil eh eh ...". It is obvious that "sos" and "sil" should not appear in an alternate order. Therefore, we implement the following rules to correct the raw alignments for TIMIT:

a. Alignments should be started with a few "sos", and followed with a few "sil";
b. Alignments should be ended with a few "sil";

We make a few modifications upon alignment method according to each rule: 1) all symbols that present before the first "sos" should be changed into "sos"; 2) all symbols that present after the first "sil" and at the same time present before the first non-"sil" symbol in the corresponding ground truth labels should be changed into that non-"sil" symbol;

For WSJ, the ground truth label for each utterance does not start and end with "sil". Their models are supervised by characters. The problem of raw alignments in WSJ mainly lies in both ends. We only need to make sure the alignments for each utterance are started with a few "sos", and the "sos" at the end due to padding should be changed into.

Note that human interventions only correct naïve mistakes for raw alignments. It may not be accurate, but it would be better for visualizing and would not be worse than the original raw alignments. After these post-processing, the alignments of each frame in the attention-based model are obtained, which will be quite helpful in visualizing t-SNE embeddings. After the t-SNE transformation and force alignment, the embeddings of the encoder outputs could be drawn, with the first and the second t-SNE components correspond to the horizontal and vertical axis.

V. ANALYZING PHONE DYNAMICS USING CCA

The SVCCA method is a powerful tool to compare two different sets of representations output from two different networks. However, in the attention-based model, we are more interested in analyzing the dynamics of localities instead of the representation of the whole utterance. It will not be easy to conclude over so many characters (including many irregular characters) without proper categorizations, therefore, we only aim for models that are supervised by phones which can be categorized by phonetics.

We divide the whole utterance into several segments. Each segment is composed of identical phones. The CCA comparison is applied between two embeddings output from two respective models. This is shown in Fig. 2. In order to avoid meaningless computation for each phone, at least the number of data points should be larger than the output dimensions. Therefore, we first implement t-SNE to reduce dimensions for object outputs. The schematic flow of this is shown in Fig. 3.

We then calculate CCA coefficients for each phone. Denote $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ as two different encoder output representations for

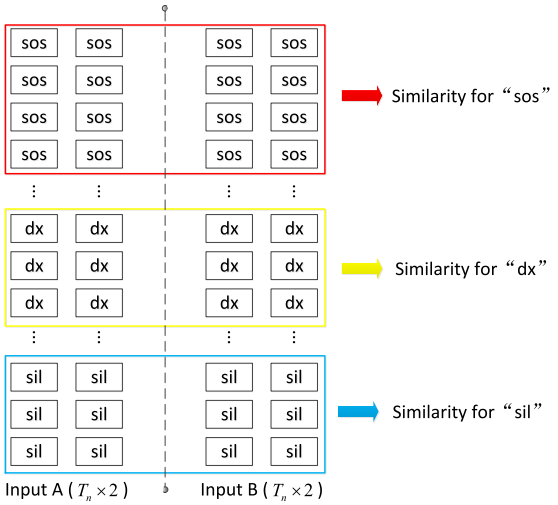


FIGURE 2. Computing similarities for phones in the utterance.

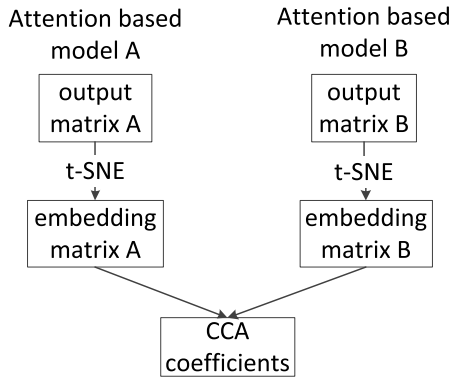


FIGURE 3. Comparisons on t-SNE embeddings using CCA.

one phone. They both have the same size of $T_n \times 2$, where T_n denotes the length of the segment. We first compute their covariance matrix $Cov(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$, and it is composed of four blocks:

$$Cov(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \begin{bmatrix} C_{\boldsymbol{Y}_1\boldsymbol{Y}_1} & C_{\boldsymbol{Y}_1\boldsymbol{Y}_2} \\ C_{\boldsymbol{Y}_2\boldsymbol{Y}_1} & C_{\boldsymbol{Y}_2\boldsymbol{Y}_2} \end{bmatrix} \tag{6}$$

We rescale each block of the matrix to get four new matrices to make CCA computation more stable:

$$\begin{cases} \boldsymbol{C}_1 = C_{\boldsymbol{Y}_1\boldsymbol{Y}_1} / \max\left(\left|C_{\boldsymbol{Y}_1\boldsymbol{Y}_1}\right|\right) \\ \boldsymbol{C}_2 = \boldsymbol{C}_3 = C_{\boldsymbol{Y}_1\boldsymbol{Y}_2} / \sqrt{\max\left(\left|C_{\boldsymbol{Y}_1\boldsymbol{Y}_1}\right|\right) \cdot \max\left(\left|C_{\boldsymbol{Y}_2\boldsymbol{Y}_2}\right|\right)} \\ \boldsymbol{C}_4 = C_{\boldsymbol{Y}_2\boldsymbol{Y}_2} / \max\left(\left|C_{\boldsymbol{Y}_2\boldsymbol{Y}_2}\right|\right) \end{cases} \tag{7}$$

Then, through SVD, the diagram matrix is obtained by:

$$\boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V} = SVD\left(\boldsymbol{C}_1^{-1} \odot \boldsymbol{C}_2 \odot \boldsymbol{C}_4^{-1}\right) \tag{8}$$

where $\cdot$ represent for the matrix multiply operation and $\odot$ is an element-wise dot operation.

Let $\boldsymbol{S}_n = (S_1, S_2)$ denote the CCA coefficients of the n_{th} segment which is composed of identical symbols. Each element represents the CCA similarity of each t-SNE embedding dimension. Denote the similarities of the n_{th} segment as C_n, it is then calculated by the mean of $\boldsymbol{S}_n$:

$$C_n = \frac{S_1 + S_2}{2} \tag{9}$$

Since we have a set of 2-dimensional vectors after t-SNE clustering, the number of continuous identical symbols in the alignments should be larger than 2. This is because CCA produces meaningless representations of symbols when the number of data points is less than the number of feature dimensions. Therefore, we make an extra modification rule for raw alignments besides rules a & b when doing CCA:

c. Each symbol should be present at least three times in a row in alignments; otherwise, they will be abandoned for CCA comparison.

Finally, the whole procedure is concluded below:

Step 1: obtain output matrix A and B from attention-based encoder A and B respectively.

Step 2: applying t-SNE to output matrix A and B, and get embedding matrix A and B.

Step 3: obtain force alignments (based on rules of a, b, and c) for both embedding matrices, and divide the whole utterances into several sequences with different identical symbols.

Step 4: compute C_n between two embedding matrices for each segment of consecutive identical symbols (as is shown in FIGURE 2).

Step 5: statistic values for all C_n, and evenly divide them into ten intervals (ranged from 0 to 1). Note that for the segments which are composed of the same alignments, they are merged for statistics. When more values are in intervals that are close to 1, it indicates that the phone converges quickly.

VI. EXPERIMENTS AND ANALYSIS

In this paper, we build models on TIMIT and WSJ. The setups for two corpora follow the settings in [30]. Utterances from the test set are selected for showing t-SNE embeddings and CCA comparisons.

A. DESCRIPTIONS OF SPEECH RECOGNITION MODELS

We first introduce details of target models and their performances. In order to better understanding the attention-based models, we choose typical models to experiment on.

For TIMIT, we choose our best-performed model in [29] of our previous work. The model is a joint CTC-attention model trained on 120-dimensional high-level features (the original 40-dimensional features with delta and delta-delta components concatenated). The high-level features are extracted through multi-lingual training and transfer learning, with dimension-reduction using convex nonnegative matrix factorization (CNMF) [34]. The multi-lingual resource we use is from Voxforge Italian, German, French, and Spanish. For the structure, the encoder has 3 BLSTM layers in the CTC part and 2 BLSTM layers in the attention part with 320 units in each layer and direction. Dropout is applied on both BLSTM and attention layers with a rate of 0.2. The attention layer

TABLE 1. Recognition results of TIMIT and WSJ.

	Methods	PER (%)
TIMIT	transferred high-level features + pure attention-based model (P0)	23.28
	transferred high-level features + joint CTC-attention (P1) [29]	16.96
WSJ	CNN + BLSTM + joint CTC-attention + word-LM (K0)	5.9
	high-level features + joint CTC-attention + RNN-LM (K1)	4.7

is location-based, with 320 units. The decoder is a 1-layer LSTM with 300 units. The system is rescored by an RNN language model (RNN-LM) with a weight of 0.2. We evaluate the recognition results by phone error rate (PER), and list them in TABLE 1.

The second model is trained with 120-dimensional high-level features, which is also concatenated with delta components as it is in TIMIT. However, what is different is that these CNMF features are extracted from DNN that is only trained with the target language. The rest settings are the same as in TIMIT.

For WSJ, we choose two different models. The first model is trained on 83 filter banks with no delta components. The encoder is a 6-layer BLSTM stacked over a CNN. There are 320 units in each direction and layer in BLSTMs, and then they are projected to 320 units. The dropout rate for the encoder is 0.2. The CNN is composed of two identical components, and each component includes consecutive two convolutional layers stacked with a max-pooling layer (the 2D max-pooling size is (3, 3)). The convolutional parameters of the CNN follow the settings in [30]. For the rest BLSTM layers, the time resolution of the second and the third BLSTM layer are both declined by half. The activation used is the rectified linear unit (ReLU). The decoder is a 1-layer LSTM with 300 units. We train a word-level RNN-LM [35] to help decode all end-to-end systems. The RNN is a 1-layer LSTM with 1000 units. The RNN-LM is trained using stochastic gradient descent (SGD) for 20 epochs with a batch size of 100. The softmax layer predicts 65000-dimensional output values, the vocabulary size is 65000.

B. VALIDATION FOR THE FORCE ALIGNMENT METHOD

Before trying to understand the encoder, we would like to prove and show why we make some specific rules to fix alignments.

1) VALIDATION IN TIMIT

We first take an utterance from the TIMIT test set as an example. The name of the utterance is "mwew0_sx11", and the ground truth label of it is "sil hh iy w el ax l aw er r eh r l ay sil". The alignments are from system P1 (see definition in Table 1, the same applies to P0, K0, and K1). The raw alignments before modifications are:

"**eh** sos sos sos sos sil sil sil sil sil sil sil sil sil sil **sos** hh hh hh hh hh hh hh iy iy iy iy iy w w w w el el el el el el el el ax ax ax ax ax ax ax ax ax ax ax ax ax l l l l l l l l l l l l l l aw aw aw aw aw aw aw aw aw er er er er er er er er er er er er er er r r r r r r r eh eh eh eh eh eh eh eh eh eh eh eh **sos sos** eh eh eh eh eh eh eh eh r r r r r l l l l l l l l l l l l l l l l ay ay ay ay ay ay ay ay ay ay ay ay ay ay sil sil sil sil sil sil sil sil sil sil sil sil".

The bolded alignments are obvious errors we have mentioned in Section IV. The alignments should be started by a few "sos" followed with some "sil", and no "sos" should be present in places other than the beginning. Then the modified alignments are:

"**sos** sos sos sos sos sil sil sil sil sil sil sil sil sil **sil** sil hh hh hh hh hh hh hh iy iy iy iy iy w w w w el el el el el el el el ax ax ax ax ax ax ax ax ax ax ax ax ax l l l l l l l l l l l l l l aw aw aw aw aw aw aw aw aw er er er er er er er er er er er er er er r r r r r r r eh eh eh eh eh eh eh eh eh eh eh eh **eh eh** eh eh eh eh eh eh eh eh r r r r r l l l l l l l l l l l l l l l l ay ay ay ay ay ay ay ay ay ay ay ay ay ay sil sil sil sil sil sil sil sil sil sil sil sil".

The bolded symbols are corrected symbols.

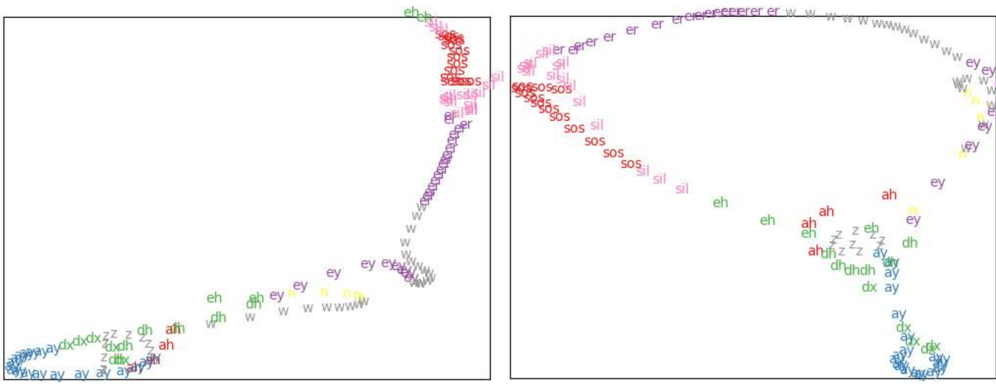
2) VALIDATION IN WSJ

We also take an utterance from WSJ eval92 as another example. The name is "444c040p" and the ground truth label of it is "T H E Y ' R E <space> J U S T <space> W A I T I N G <space> F O R <space> T H E <space> O T H E R <space> S H O E". We achieve the raw alignments from system K1:

"**T T** sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos T T T T T T T T T T T T T T T H H H H E Y ' R R E E E <space><space> J J J J J U U U U S S S S S T T T T <space><space><space><space> W W W W W A A A A A I I I T T T I I I I I N N N G G G G <space><space> <space><space><space> F F F O O R R R <space><space> T T T H H E E <space><space><space><space> O O O O T T T T H H H E E E R R R <space><space><space><space> <space><space> S S S S S S H H H H H O O O O O O O O O O O O E **T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T T**".

We also bold obvious errors for raw alignments. The alignments should be started by a few "sos" followed with some "sil", and no "sos" should be present in places other than the beginning. Then we bold the corrected symbols in the modified alignments:

"**sos sos** sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos



(a) (b)

FIGURE 4. **T-SNE on TIMIT utterance “mjdh0_si1984” for P0 (a) and P1 (b). Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.**

sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos sos T T T T T T T T T T T T T T T H H H H E Y ’ R R E E E <space><space> J J J J J J U U U U S S S S S T T T T <space><space><space><space> W W W W W A A A A A I I I T T T I I I I I N N N G G G G <space><space> <space><space><space> F F F O O R R R <space><space> T T T H H E E <space><space><space><space> O O O O T T T T H H H E E E R R R <space><space><space><space> <space><space> S S S S S S H H H H H O O O O O O O O O O O O E **E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E**”.

We correct these alignments to avoid as many errors as possible so that it would be more accurate to find patterns based on t-SNE figures. After modification, the location of each symbol in the alignments exactly follows the order of the symbols in the ground truth label.

C. VISUALIZING THE ENCODER OUTPUTS

We first take outputs from the last layer of the encoder. The output vectors are 320-dimensional, corresponding to the number of the units in the output project layer. Before t-SNE, we first use K-means for initialization. The perplexity is 30 for t-SNE, which is an empirical setting. For each corpus, we use alignments obtained from the best-performed model, model P1. This is because it is most likely able to provide the most accurate attention weights.

1) COMPARISONS BETWEEN THE ATTENTION-BASED MODEL AND THE JOINT CTC-ATTENTION MODEL

To compare the attention-based model and the joint CTC-attention model, we first take two utterances from the TIMIT test set. The first one is utterance “mjdh0_si1984”, which is the shortest utterance in the set. Its ground truth label is “sil eh n w ah dx ay z dh ey w er sil”. The other one is a much longer utterance “mjdh0_sx274”, and the label is “sil k l ih f w ix s uw dh vcl b ay dh ax l ix vcl zh er r iy ix s epi m ix s aa zh sil”. We draw t-SNE embeddings of “mjdh0_si1984” and “mjdh0_sx274” in Fig. 4 and Fig. 5 separately.

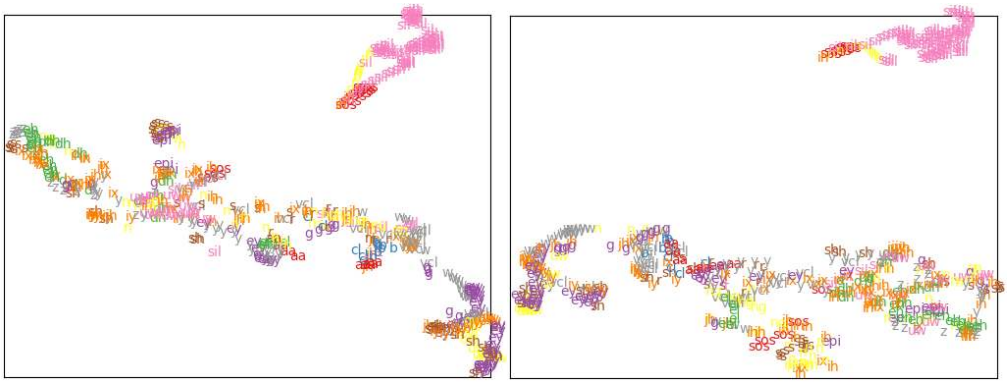
From Fig. 4 (a), we can see that model P0 fail to cluster continuous identical alignments. For example, the symbol “eh” only appears once in the ground truth label, while the alignments of “eh” are in separate areas.

However, Fig. 4 (b) shows that P1 separates different symbols by into manifolds. The aligned embeddings are basically shaped into a circle in sequential order, started by “sos…sill…eh…” and ended by “w…er…sil”.

However, Fig. 5 shows that both model P0 and P1 could not explicitly separate different symbols for a long utterance. Many different alignments overlap in similar areas and the manifolds are not clear. But still, the joint model P1 could separate symbols between “dh” and “eh”, while P0 could not. This also explains why attention-based models have difficulties decoding long utterances: the encoder could not explicitly distinguish different symbols at different locations, and this leads to vague attention weights to some utterances.

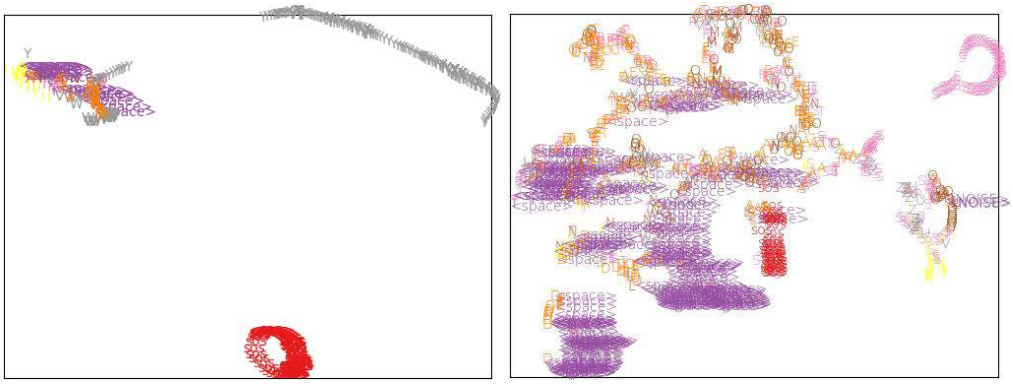
We further experiment on WSJ, Fig. 6 (a) & (b) show t-SNE results for a short utterance “440c040j” and a long utterance “447c040d” in model K1, and Fig. 7 (a) & (b) show t-SNE results for a short utterance “440c040j” and a long utterance “447c040d” in model K0. Their ground truth labels are “I T <space> W A S N ’ T <space> A <space> G I V E A W A Y” and “<NOISE><space> A L S O <space> M E N T I O N E D <space> W A S <space> A <space> C O N T R O V E R S I A L <space> P R O P O S A L <space> T O <space> D E N Y <space> T H E <space> D E D U C T I O N <space> F O R <space> T W E N T Y <space> P E R C E N T <space> O F <space> C O R P O R A T E <space> A D V E R T I S I N G: <space> C O S T S <space> A N D <space> T O <space> R E Q U I R E <space> I N S T E A D <space> T H A T <space> T H E Y <space> B E <space> A M O R T I Z E D <space> O V E R <space> T W O <space> Y E A R S” respectively.

Fig. 6 (a) and Fig. 7 (a) show good manifolds of the encoder outputs, and it successfully separates the starting



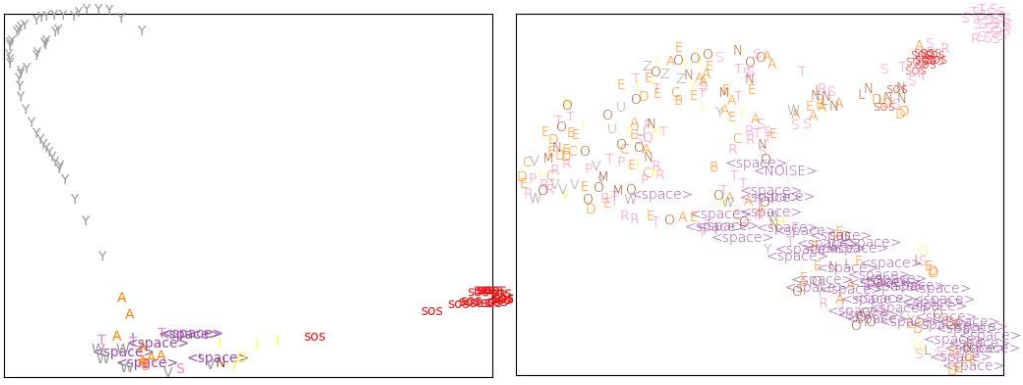
(a) (b)

FIGURE 5. T-SNE on TIMIT utterance "mjdh0_sx274" for P0 (a) and P1 (b). Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.



(a) (b)

FIGURE 6. T-SNE on WSJ utterance "440c040j" (a) and "447c040d" (b) for K1. Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.



(a) (b)

FIGURE 7. T-SNE on utterance "440c040j" (a) and "447c040d" (b) for K0. Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.

symbol and the ending symbol. However, the embeddings of the long utterance in Fig. 6 (b) and Fig. 7 (b) are not properly divided. For example, the first symbol "<NOISE>" is distantly located to the start symbol "sos", and is closely located to irrelevant symbols like "O" and "V". This reflects that the output vectors are not well sequentially connected.

2) COMPARISONS AMONG DIFFERENT NUMBER OF UTTERANCES

Next, we experiment with different number of utterances. We have already known what one utterance is distributed visually using t-SNE. We also would like to know what patterns could be found when embeddings from multiple utterances are shown together.

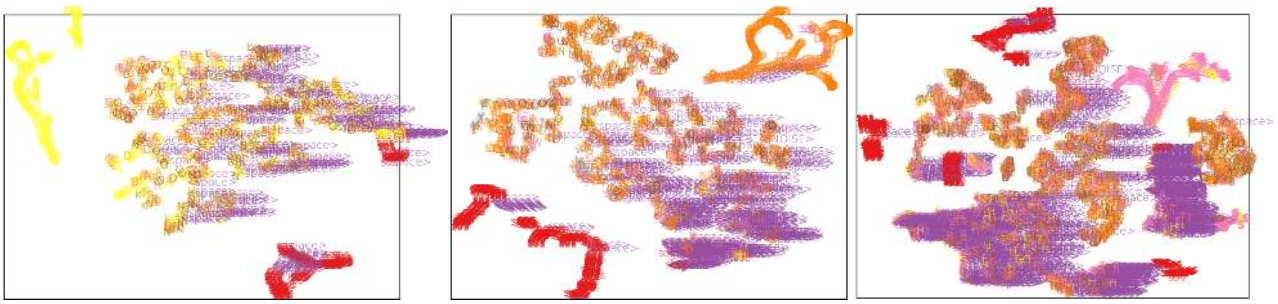


FIGURE 8. T-SNE on 3 utterances (a) & 5 utterances (b) & 10 utterances (c) for K0. Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.

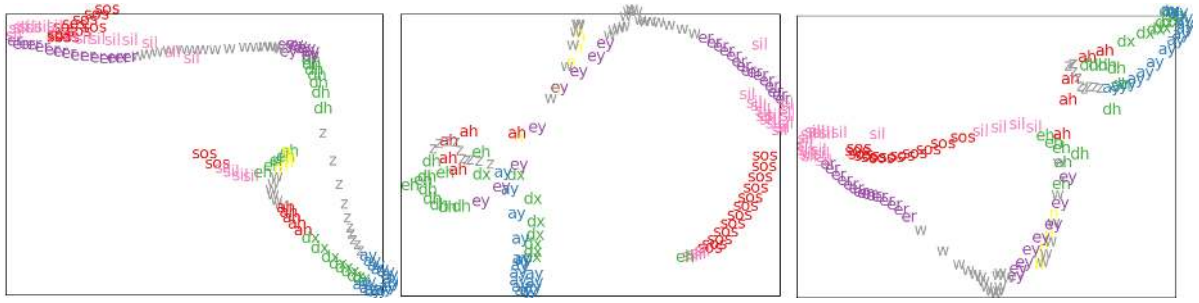


FIGURE 9. T-SNE on "mjdh0_si1984" for P1 at epoch 4 (a) & 8 (b) & final epoch (c). Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.

TABLE 2. utterance list of WSJ.

Number of utterances	Names of utterances
3	441c040i, 444c040c, 447c040a
5	441c040i, 442c0411, 444c040c,446c040k, 447c040a
10	441c040i, 442c0407, 442c0411,443c040q, 444c040c, 444c0416,445c040u, 446c040k, 447c040a,447c0415

We respectively choose 3, 5, and 10 utterances from WSJ eval92 set and draw their aligned t-SNE embeddings in Fig. 8. The names of chosen utterances are listed in TABLE 2. Although it is impossible to separate one manifold from another in the figure, the embeddings with the same alignment symbols are basically located in the same area. The symbols may be hard to recognize visually due to a limited size and the overlapping problem, but the gathering of the same colored symbols indicate that the encoder of the attention-based model is well trained for temporal modeling.

3) COMPARISONS BETWEEN DIFFERENT ENCODERS

We also compare between encoders of K0 and K1. Note that the number of the output vectors for K0 is only one-fourth of the number of the output vectors for K1 due to the time resolution reduction. However, we get similar conclusions for K0. The embeddings of the short utterance still clearly form a manifold of sequential symbols in the ground truth label, while the embeddings of the long utterance are not shaped into regular manifolds. This could explain why time resolution reduction works for the attention-based model: many neighbored frames share the same alignment, and keeping one frame out of every two frames still enables them to form a complete utterance.

4) COMPARISONS AMONG DIFFERENT EPOCHS

We further would like to know the differences among different training stages for the attention-based model using our methods. We use the same utterances in TIMIT as we used in previous experiments. Fig. 9 (a) & (b) & (c) respectively show t-SNE results on P1 at epoch 4, epoch 8, and the final epoch for the short utterance "mjdh0_si1984". Fig. 10(a) & (b) & (c) respectively show t-SNE results on P1 at epoch 4, epoch 8, and the final epoch for a longer utterance "mjdh0_sx274".

For the short utterance "mjdh0_si1984", the aligned embeddings show that the encoder forms better and better manifolds of the sequence with the training goes on. In Fig. 9(a), both "sos" and "sil" are distantly separated, which should be located together. In Fig. 9(b), "eh" should follow "sos" and "sil", instead they are located far away. From the results at the final epoch of the model, none of the above problems exists. The aligned

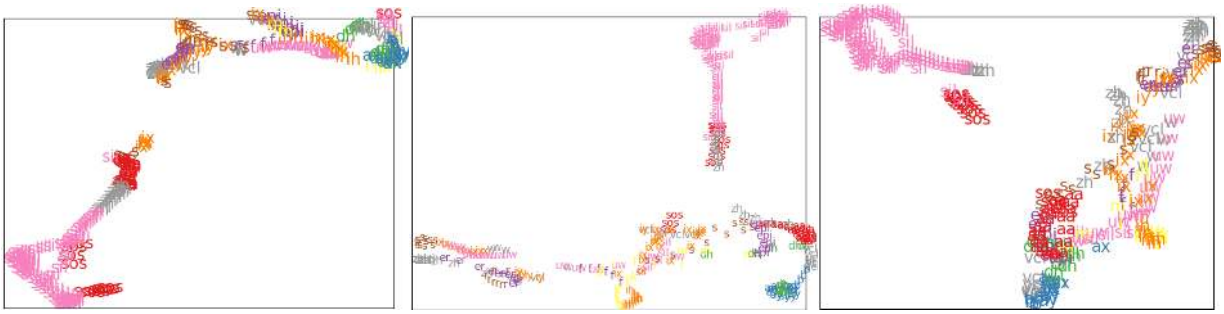


FIGURE 10. T-SNE on "mjdh0_sx274" for P1 at epoch 4 (a) & 8 (b) & final epoch (c). Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.

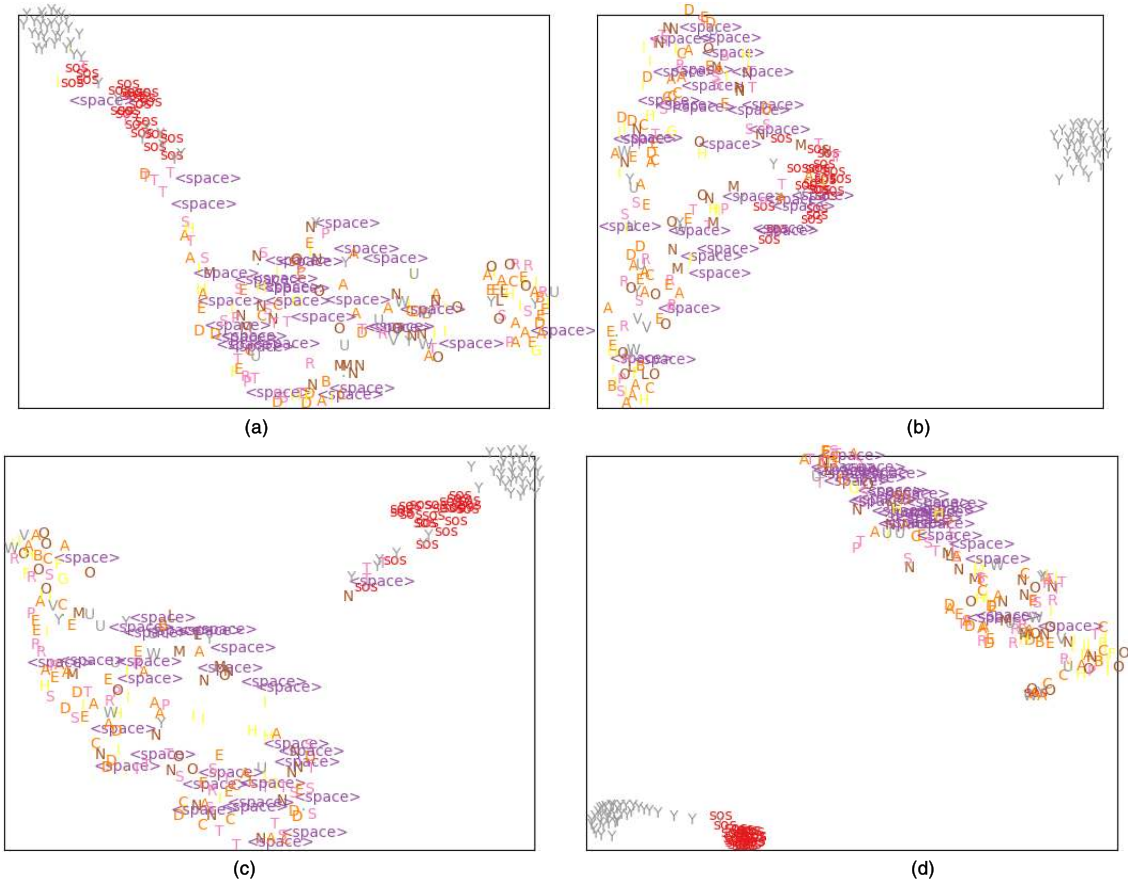


FIGURE 11. T-SNE on "444c040c" of the CNN outputs (a) & the first BLSTM layer outputs (b) & the third BLSTM layer outputs (c) & the last BLSTM layer outputs (d) in K0. Horizontal axis: the 1_{st} dimension of t-SNE embeddings; vertical axis: the 2_{nd} dimension of t-SNE embeddings.

embeddings show relative better manifolds which start from "sos……sil……eh……", and end at "……er……sil".

For the longer utterance "mjdh0_sx274", the case is not the same with the short utterance. We could see that embeddings are not sequentially ranged in all three sub-figures. In Fig. 10 (a), "sos" symbols are separated. In both Fig. 10 (b) and 9 (c), the beginning symbols are not well ranged and they are overlapped with other symbols, even with ending symbols like "zh". Besides, "sos" symbols are distantly separated in both Fig. 10 (a) and 10 (c). Since these show the dynamics of the encoder outputs, it can be concluded that the improvements are less for a relative long utterance than a short utterance during training.

5) COMPARISONS AMONG OUTPUTS OF DIFFERENT LAYERS

At last, we study outputs of different layers in the encoder, and show them in Fig. 11. We choose model K0 to experiment on since it has more layers than other models in our experiments. The model K0 is composed of a CNN and a 6-layer BLSTM. We compare outputs from the last layer of CNN in Fig. 11(a),

TABLE 3. Ratios of cca coefficients values among different phone categories for P0.

category	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Plosive	0	0	0	0.01	0.02	0	0.15	0.19	0.24	0.39
Strong fricative	0	0	0	0	0.03	0	0.21	0.21	0.24	0.31
Weak fricative	0	0	0	0	0.02	0	0.18	0.19	0.26	0.35
Nasal/ Flap	0	0	0	0	0.02	0	0.20	0.19	0.25	0.34
Semi-vowel	0	0	0	0	0.03	0	0.17	0.22	0.27	0.30
Short vowel	0	0	0	0	0.02	0	0.21	0.23	0.24	0.30
Long vowel	0	0	0	0	0.02	0	0.21	0.24	0.25	0.26
Silence	0	0	0	0.01	0.02	0	0.19	0.20	0.27	0.32

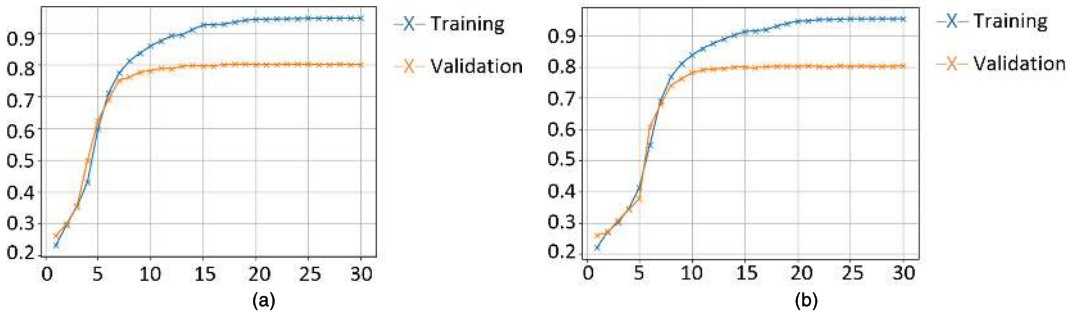


FIGURE 12. Training accuracy of the model P0 (a) and P1 (b). Horizontal axis: the training epochs; vertical axis: the training accuracy.

the first BLSTM layer in Fig. 11(b), the third BLSTM layer in Fig. 11(c), and the last BLSTM layer in Fig. 11(d). We take the utterance "444c040c" as an example to show t-SNE results. Its ground truth label is "I N <space> A <space> S E P A R A T E <space> I N C I D E N T <space> A <space> U. <space> S. <space> D E S T R O Y E R <space> F I R E D <space> M A C H I N E <space> G U N <space> W A R N I N G <space> S H O T S <space> A T <space> T W O <space> S M A L L <space> U N I D E N T I F I E D <space> B O A T S <space> T H A T <space> A P P R O A C H E D <space> T H E <space> C O N V O Y".

First, we can see that the t-SNE embeddings of CNN outputs are quite different than BLSTM outputs. The aligned embeddings of CNN outputs are not fully spatially separated in Fig. 11(a).

Especially, both the starting symbol "sos" and the ending symbol "Y" are overlapped with many other symbols in similar locations. For BLSTM layers, the spatial distribution is more compact when the outputs are from a deeper layer. The same aligned embeddings from the deeper BLSTM layer are located more closely.

Overall, if the encoder embeddings are not clearly shaped into manifolds of the ground truth label sequences, the model would not have a good performance due to bad attentions. Also, the embeddings that output from a CNN or a shallow BLSTM layer could not produce clear manifolds indicate that a certain number of recurrent layers are necessary for the attention-based model.

D. ANALYSIS OF DYNAMICS OF PHONES

In this part, we experiment on the training dynamics of the encoder using t-SNE and CCA. We experiment on TIMIT since it is modeled with phones.

According to the method in chapter V, we first compute CCA coefficients for each phone in each utterance. The alignments are all obtained from the best-performed model (model P1). The comparison is carried out between the epoch 4 and the final epoch. Fig. 12(a) & (b) respectively show the accuracy of model P0 & P1 in different training stages. Epoch 4 is chosen as a representative for its low accuracy during the early training stage.

We then statistic their values following the procedure in section V, step 5. We respectively experiment on model P0 and P1 and summarize their ratios in TABLE 3 and TABLE 4. We summarize the phone labels into phone categories according to TABLE 5, which is based on [36] and [37]. If the ratio that represents the higher domain of values is large, it means that the phone category tends to converge in an early stage during training.

Similar results are observed from both models. Almost 40% of the plosive and the nasal/flap embeddings vary in a limited scale after epoch 4. The strong fricative, weak fricative, semi-vowel, and short vowel have almost the same distributions for ratios of CCA coefficient values, indicating that their dynamics are similar through the whole training period. The long vowel does not have large ratios for higher domains comparing with other phone categories. They converge slowly during training.

TABLE 4. Ratios of CCA coefficients values among different phone categories for P1.

category	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Plosive	0	0	0	0	0.01	0	0.18	0.17	0.25	0.39
Strong fricative	0	0	0	0	0.01	0	0.18	0.20	0.30	0.31
Weak fricative	0	0	0	0	0.01	0	0.19	0.21	0.27	0.32
Nasal/ Flap	0	0	0	0	0.01	0	0.16	0.20	0.25	0.38
Semi-vowel	0	0	0	0	0	0	0.20	0.20	0.29	0.31
Short vowel	0	0	0	0	0.02	0	0.19	0.21	0.25	0.32
Long vowel	0	0	0	0	0.02	0	0.22	0.25	0.26	0.24
Silence	0	0	0	0.01	0.02	0	0.17	0.20	0.25	0.34

TABLE 5. Phone categorization.

category	phone labels
Plosive	g, d, b, k, t, p
Strong fricative	s, z, sh, zh, ch, jh
Weak fricative	f, v, th, dh, hh
Nasal/Flap	m, n, en, ng, dx
Semi-vowel	l, el, r, w, y
Short vowel	ih, ix, ae, ah, ax, eh, uh, aa
Long vowel	iy, uw, ao, er, ey, ay, oy, aw, ow
Silence	sil, epi, q, vcl, cl

The fast-convergent phone such as the plosive or the nasal/flap belongs to burst voices, while the slow convergent phone like the long vowel has a long vocal duration, and has tonal slippage. As is known, the attention-based model is difficult at decoding long sentences [38]–[40], we further demonstrate that the attention-based model is also difficult at modeling phones with long tones in the speech recognition.

However, what should be emphasized is that the convergent speed does not necessarily have a causal correlation to the classification accuracy of that phone category. The conclusions drawn from our CCA experiment are only helpful for understanding the training dynamics of phones, and are potentially useful for further improvements upon the attention-based model.

VII. CONCLUSION

In order to better understand the attention-based model, we propose to use a human-intervened force alignment method to align the t-SNE embeddings of the encoder outputs. In addition, we propose a method of combining t-SNE and CCA to analyze the training dynamics of phone categories in the attention-based model. Examples from both TIMIT and WSJ validate the necessity of human interventions during aligning. The aligned embeddings of the encoder outputs are shaped into manifolds of the ground truth label sequences visually, demonstrating the effectiveness of the attention mechanism in speech recognition. Outputs that come from the deeper layer, the shorter utterance, or the better-performed models in speech recognition, tend to produce clearer manifolds. Besides, embeddings with the same aligned symbols tend to gather at similar positions when multiple utterances are drawn together. This proves the consistency of our method. Further experiments on analyzing the phone dynamics show that the long vowel phone tends to converge slowly, while the plosive and the nasal/flap phone converge quickly.

REFERENCES

[1] J. Picone, "Continuous speech recognition using hidden Markov models," *IEEE ASSP Mag.*, vol. 7, no. 3, pp. 26–41, Jul. 1990.
[2] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, "A review on speech recognition technique," *Int. J. Comput. Appl.*, vol. 10, no. 3, pp. 16–24, Nov. 2010.
[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.
[4] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173–182.
[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 577–585.
[6] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1764–1772.
[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.
[8] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 518–529.
[9] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*. [Online]. Available: https://arxiv.org/abs/1308.0850
[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd ICLR*, 2015, pp. 1–15. [Online]. Available: https://arxiv.org/abs/1409.0473
[11] K. Xu, J. Ba, and R. Kiros, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[13] N. Jaitly, D. Sussillo, Q. V. Le, O. Vinyals, I. Sutskever, and S. Bengio, "A neural transducer," 2015, *arXiv:1511.04868*. [Online]. Available: https://arxiv.org/abs/1511.04868
[14] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 206–213.
[15] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.

[16] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," 2016, *arXiv:1612.02695*. [Online]. Available: https://arxiv.org/abs/1612.02695
[17] N. U. Islam and S. Lee, "Interpretation of deep CNN based on learning feature reconstruction with feedback weights," *IEEE Access*, vol. 7, pp. 25195–25208, 2019.
[18] C.-C.-J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 406–413, Nov. 2016.
[19] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with SincNet," 2018, *arXiv:1811.09725*. [Online]. Available: https://arxiv.org/abs/1811.09725
[20] L. Bai, P. Weber, P. Jancovic, and M. Russell, "Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features," in *Proc. INTERSPEECH*, Aug. 2018, pp. 1472–1476.
[21] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, and M. Delcroix, "Semi-supervised end-to-end speech recognition," in *Proc. INTERSPEECH*, Aug. 2018, pp. 2–6.
[22] J. Kim, G. Englebienne, and K. P. Truong, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," 2017, *arXiv:1708.03920*. [Online]. Available: https://arxiv.org/abs/1708.03920
[23] Z. Tang, Y. Shi, D. Wang, Y. Feng, and S. Zhang, "Memory visualization for gated recurrent neural networks in speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2736–2740.
[24] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6076–6085.
[25] A. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," *Advances in Neural Information Processing Systems*, 2018, pp. 5727–5736.
[26] B. J. Hou and Z. H. Zhou, "Learning with Interpretable Structure from RNN," 2018, *arXiv:1810.10708*. [Online]. Available: https://arxiv.org/abs/1810.10708
[27] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.
[28] T. Hori, S. Watanabe, and J. Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 1–14.
[29] C.-X. Qin, D. Qu, and L.-H. Zhang, "Towards end-to-end speech recognition with transfer learning," *EURASIP J. Audio, Speech, Music Process.*, vol. 2018, p. 18, Nov. 2018.
[30] C. Qin, W. Zhang, and D. Qu, "A new joint CTC-attention-based speech recognition model with multi-level multi-head attention," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, p. 18, Dec. 2019.
[31] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4845–4849.
[32] T. Hori, S. Watanabe, and Y. Zhang, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," 2017, *arXiv:1706.02737*. [Online]. Available: https://arxiv.org/abs/1706.02737
[33] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[34] C. Ding, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.
[35] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," 2018, *arXiv:1808.02608*. [Online]. Available: https://arxiv.org/abs/1808.02608
[36] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification 1," in *Proc. EUROSPEECH*, Berlin, Germany, 1997, pp. 401–404.
[37] H. Huang, Y. Liu, L. Ten Bosch, B. Cranen, and L. Boves, "Locally learning heterogeneous manifolds for phonetic classification," *Comput. Speech Lang.*, vol. 38, pp. 28–45, Jul. 2016.
[38] W. Cai, Z. Cai, W. Liu, X. Wang, and M. Li, "Insights in-to-end learning scheme for language identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5209–5213.
[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473
[40] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2016, pp. 207–212.

CHU-XIONG QIN received the B.S. and M.S. degrees in information and communication from the National Digital Switching System Engineering and Technological Research and Development Center, Zhengzhou, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in speech recognition with the PLA Strategic Support Force Information Engineering University. His research interests are in speech signal processing, continuous speech recognition, and machine learning.

DAN QU received the M.S. degree in communication and information system from the Xi'an Information Science and Technology Institute, Xi'an, China, in 2000, and the Ph.D. degree in information and communication engineering from the National Digital Switching System Engineering and Technological Research and Development Center, Zhengzhou, China, in 2005. She is currently a Professor with the PLA Strategic Support Force Information Engineering University. Her research interests are in speech signal processing and pattern recognition.