Speaker Adaptive Training for Speech Recognition Based on Attention-over-Attention Mechanism

Genshun Wan[1]*,Jia Pan*[1]*,Qingran Wang*[2]*,Jianqing Gao*[2]*,Zhongfu Ye*[1]

[1]University of Science and Technology of China
[2]iFlytek Research, iFlytek Co., Ltd.

{gswan,panjia}@mail.ustc.edu.cn, {qrwang2,jqgao}@iflytek.com, yezf@ustc.edu.cn

Abstract

In our previous work, we introduced a speaker adaptive training method based on frame-level attention mechanism for speech recognition, which has been proved an effective way to do speaker adaptive training. In this paper, we present an improved method by introducing the attention-over-attention mechanism. This attention module is used to further measure the contribution of each frame to the speaker embeddings in an utterance, and then generate an utterance-level speaker embedding to perform speaker adaptive training. Compared with the frame-level ones, the generated utterance-level speaker embeddings are more representative and stable. Experiments on both the Switchboard and AISHELL-2 tasks show that our method can achieve a relative word error rate reduction of approximately 8.0% compared with the speaker independent model, and over 6.0% compared with the traditional utterance-level d-vector-based speaker adaptive training method.

Index Terms: speech recognition, speaker adaptive training, attention-over-attention

1. Introduction

Recently, deep neural networks (DNNs) such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have become the mainstream structure of automatic speech recognition (ASR) [1–3]. However, when it comes to the speakers with special accents or pronunciation habits, the accuracy of ASR may suffer a significant reduction. In response, speaker adaptive training (SAT) is one of the effective approaches to improve the performance of ASR on these conditions.

The most widely used methods of SAT can be classified into two categories: auxiliary features and adversarial learning. Auxiliary features that contain information about speakers are used to perform speaker adaptive training. Speaker i-vectors or bottleneck vectors, obtained by a pretrained speaker recognition model, can be used with the acoustic features together to make the acoustic model generalize better to different speakers [4–7]. In addition, speaker codes [8–10] can also be used to represent speaker characteristics. To highlight the importance of the speaker embeddings in adaptation, the authors in [11, 12] try to generate the speaker-dependent (SD) parameters via a controller network that takes speaker embeddings as input, and the controller network is shared among all speakers. Another type of methods to perform SAT is using the adversarial learning scheme. Similar to the methods used in domain adaptation [13–15], the acoustic model and the speaker classification model are jointly optimized via adversarial learning [16]. In [17], a reconstruction network is trained to predict the input speaker i-vector. The mean-squared error loss of the i-vector reconstruction and the cross-entropy loss of the acoustic model are jointly optimized through adversarial multi-task learning.

The speaker adaptive training methods mentioned above should be helpful when a number of adaptation data is provided. However, in real-world ASR systems, the collection of sufficient speaker data is very difficult, particularly the labeled data. Insufficient data will introduce an inaccurate speaker embedding and make a sharp decline in performance for speaker adaptive training. As one of the mainstream approaches, the i-vector-based speaker adaptive training method takes the i-vectors obtained in advance as the speaker embeddings. However, their performance is unsatisfactory because the i-vector is obtained without regard to the speech recognition task.

In order to provide a dynamic speaker embedding associated with the speech recognition performance, a speaker adaptive training method based on attention mechanism is proposed in our previous work [18]. The i-vectors of all speakers in the training data are obtained as a static memory in advance. For each frame, the closest speaker i-vector is selected with attention mechanism which is learned jointly with the acoustic model from the training data. However, subject to the limited and partially invalid frames, the speaker representation would be unstable and uniform to a certain extent.

In this paper, we propose a speaker adaptive training method for speech recognition based on attention-over-attention mechanism. For each utterance, the nearest d-vectors are selected and then recombined to an utterance-level aggregated vector by attention-over-attention mechanism. The aggregated vector is connected with the acoustic model to provide the information about the current speaker. Compared with the traditional utterance-level d-vectors or the frame-level aggregated d-vectors metioned in [18], aggregated utterance-level vectors can provide a more accurate and robust speaker representation to improve the recognition accuracy. Experiments on the Switchboard and AISHELL-2 tasks show that the proposed method achieves a significant improvement over the SD method based on utterance-level d-vectors.

2. Related Work

Since the speaker representation is essential to the speaker adaptive training, there have been intensive researches to optimize the speaker embeddings and create the direct relationship between the speaker embeddings and the speech recognition performance with limited resources. As mentioned above, an attention based speaker adaptive training method is proposed in [18]. As illustrated in Fig. 1, the framework mainly consists of two parts: the main network and the attention module.

The main network of the proposed method is the same as other structures of acoustic model, including feedforward neural networks, CNNs and RNNs. The main network plays two roles: acoustic modeling and providing information for the attention module. As a kind of weak information, speak char-

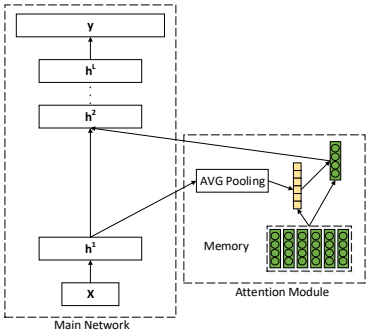

Figure 1: *The framework of the speaker adaptive training method based on attention mechanism.*

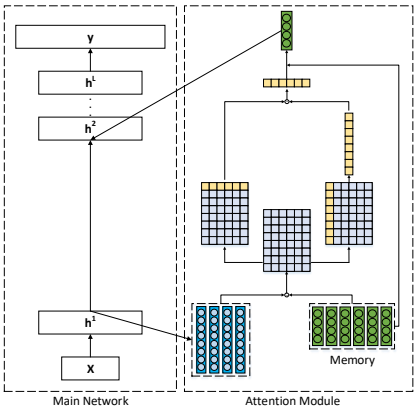

Figure 2: *The framework of the speaker adaptive training method based on attention-over-attention mechanism.*

acteristics need to be extracted and used before it is removed finally by deep neural networks. Therefore, the outputs of the hidden layers near the input layer are provided for the attention module.

The attention module is equipped mainly to select the vectors that are most similar to the current frame from the memory and combine them into a vector named the frame-level aggregated speaker vector. The memory consists of a group of vectors, such as i-vectors [19] and d-vectors [20], which are easily distinguished from each other by its corresponding speaker. As an effective representation of the speaker, the aggregated speaker vector based on attention mechanism is used together with the acoustic features to do speaker adaptive training.

Experiments on the Switchboard task show that the speaker adaptive training method based on attention mechanism could achieve a decent performance improvement compared to that of the i-vector-based speaker adaptive training method.

3. The Proposed Method

3.1. Motivation

In our previous work [18], the frame-level aggregated speaker vectors based on attention mechanism is used to represent the frame-level speaker embeddings. However, because only the history part of current utterance can be used to gather the speaker information during the process of attention module at each frame, especially for the first few frames, the speaker representation would be unstable and uniform. In addition, during the process of gathering the speaker information, an average pooling is used to obtain the information. Average pooling means all the history frames have the same importance. When there are some abnormal frames with little speaker information, such as silence or environmental noise frame, average pooling strategy is obviously unreasonable to generate the representative speaker embeddings. That is to say, effective speaker information may be further weaken by the partially invalid frames. In order to make better use of the long-term speaker information and then form a representative utterance-level speaker embedding, the attention mechanism should focus not only on the importance distribution of each vector in the memory at each frame, but also the importance distribution of each frame. In order to maintain the coherence and consistency of the speaker embeddings within an utterance, we tend to make use of the utterance-level embeddings with attention mechanism.

Attention mechanism is widely used in many fields, such as machine translation and speech recognition. By putting different weights on different types of information, the process of model training becomes more flexible. Common attention mechanisms include: soft attention [21], hard attention [22] and self-attention [23]. In paper [24], another attention mechanism is placed over the existing attention to further strengthen the importance of each individual attention part. This kind of attention mechanism covering two dimensional space, called attention-over-attention mechanisms, offers a potential path to generate robust utterance-level embeddings. Further attention in time dimension can weaken the influence of some abnormal frame-level speaker embedding. Therefore, the embedding generated from all the frames of this utterance can be uniform and stable.

Starting with the generation of representative and robust utterance-level speaker embeddings, we propose a speaker adaptive training method for speech recognition based on attention-over-attention mechanism

3.2. SAT Based on Attention-over-Attention Mechanism

As illustrated in Fig. 2, the main structure of the speaker adaptive training method based on attention-over-attention mechanism is similar with our previous work. We replace the common attention mechanism by attention-over-attention mechanism to generate a more representative and stable speaker embedding.

In our study, the d-vectors of the speakers in training set are extracted as the memory. To obtain the d-vectors, a neural network is pre-trained by speaker discriminative criteria such as cross-entropy or triplet loss. Then, the output of the last hidden layer is obtained to produce a frame-level speaker representation, and all the frame-level representations are then averaged to form an utterance-level speaker embedding called the d-vector. Finally, the simplified method of clustering such as K-means is adopted to reduce the number of base vectors in a memory. Assuming that the memory has N vectors, the memory is denoted by $\mathbf{m} = \{\mathbf{m}_1,\mathbf{m}_2,...,\mathbf{m}_N\}$, in which $\mathbf{m}_i$ represents the i-th vector in this memory.

Given an utterance with T speech frames, the acoustic features of the main network are represented by $\mathbf{X} = \{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T\}$, where $\mathbf{x}_t$ represents the feature vector at the frame t. The corresponding outputs of the l-th hidden layer of the main network are denoted as $\mathbf{H}^l = \{\mathbf{h}_1^l,\mathbf{h}_2^l,...,\mathbf{h}_T^l\}$.

After obtaining the output of the hidden layer near the input layer $\mathbf{h}^{low}$ and the memory $\mathbf{m}$, we calculate a similarity degree matrix, which indicates the similarity scores between speaker information for each frame and each vector in the memory. We compute the matrix $\mathbf{M} \in \mathbb{R}^{|T|\times|N|}$ by the dot product between

the transformation output vector of $\mathbf{h}^{low}$ at the frame t and the i-th memory vector.

$$M(t,i) = (\mathbf{W}^m \mathbf{h}_t^{low}) \odot \mathbf{m}_i^\top \tag{1}$$

Based on the similarity degree matrix $\mathbf{M}$, we apply a row-wise softmax function to get the similarity scores for each row. We denote $\alpha(t) \in \mathbb{R}^{|N|}$ as the memory-level attention at the frame t. $\alpha(t)$ indicates the similarity degree to each vector at this frame, as described in the following formula.

$$\begin{aligned} \alpha(t) &= softmax(M(t,1), ..., M(t,|N|)) \\ \alpha &= [\alpha(1), \alpha(2), ..., \alpha(|T|)] \end{aligned} \tag{2}$$

The contributions to the speaker embeddings of each frame are obvious different, particularly those featured as environmental noise, and so on. On the contrary, the utterance-level speaker embeddings are more robust than the frame-level ones. Instead of averaging α at all the frames to form a final attention score, another attention mechanism is introduced to determine the importance of each individual attention.

We first calculate a column-wise softmax attention to get the similarity scores for each column, and denote $\beta(i) \in \mathbb{R}^{|T|}$ as the frame-level attention at the memory i. $\beta(i)$ indicates the importance degree of each frame corresponding to the i-th vector in the memory, as described in the following formula.

$$\beta(i) = softmax(M(1,i), ..., M(|T|, i)) \tag{3}$$

Then, we average all the $\beta(i)$ to get an averaged attention β. β is also an attention score vector with T dimensions, and it can be taken as the final importance degree of each frame.

$$\beta = \frac{1}{N} \sum_{i=1}^{N} \beta(i) \tag{4}$$

Since α is a combination of memory-level attention from 1 to T and β is a frame-level attention, we can calculate the matrix multiplication between α and β to get the final attention score $\mathbf{a}$ in an utterance. Attention score $\mathbf{a}$ is a N-dimensional vector.

$$\mathbf{a} = \alpha \odot \beta^\top \tag{5}$$

The normalized attention values $\mathbf{a}$ are used to compute a weighted sum of the vectors in the memory and formula the final utterance-level aggregated speaker vector $\mathbf{c}$.

$$\mathbf{c} = \sum_{i=1}^{N} \alpha_i \mathbf{m}_i \tag{6}$$

Finally, we connect the aggregated speaker vector with the main network, and the main network generates speaker-normalized representations using the speaker information in the aggregated speaker vector. For each frame, the aggregated speaker vector $\mathbf{c}_t$ is the same as $\mathbf{c}$. A simple connection method is concatenating the aggregated speaker vector with the outputs of the hidden layers of the main network as $\hat{\mathbf{h}}_t^{low} = [\mathbf{h}_t^{low\top}, \mathbf{c}_t{}^\top]^\top$.

The final loss function at the cross-entropy training stage of the proposed method is described by the following formula:

$$\mathcal{L} = \sum_{s=1}^{S} \sum_{t=1}^{T_s} \log p(y_t^s | \mathbf{x}_t^s, \mathbf{m}) \tag{7}$$

In Eq.(7), $\mathbf{x}_t^s$ and y_t^s indicate the acoustic feature vector and the triphone state label for frame t in utterance s. T_s is the number of frames of utterance s, and S is the number of total utterances in the training set.

4. Experiments and result analysis

4.1. Experimental Setup

We evaluated the performance of the proposed approach on both English and Mandarin speech recognition tasks.

The English training data of the Switchboard (SWB) task [25] consists of 20-hour English CALLHOME and 309-hour Switchboard-I dataset, including a total of 5110 speakers. The SWB part of NIST 2000 Hub5 evaluation set is taken as test set, and it contains 1831 utterances from 40 speakers in total. The Mandarin training data of AISHELL-2 task [26] consists of 1000 hours of clean audio segments recorded via the iPhone channel from 1991 speakers, including 1293 speakers with slight northern accents, 678 speakers with southern accents and 20 speakers with other accents. The test set contains 5000 utterances from 10 speakers, and each speaker has approximately half an hour of audio segments.

4.2. Baseline systems

The SI baseline was trained with a VGG-like [27] model architecture based on frame-level cross-entropy criterion. The inputs of the model were the 40-dimensional log Mel-scale filter-bank features. The architecture of the model mainly consisted of convolutional and pooling layers, and each convolutional layer was equipped with a standard ReLU activation function. We shuffled the utterances in training data and grouped them into mini-batches with a limit of 2048 frames per minibatch to speed up training. Stochastic gradient descent was used as the optimizer, and the initial learning rate was set to 0.02. All subsequent experiments were performed by the CAFFE toolkit [28] and run on a server equipped with 4 Tesla P40 GPUs.

In our paper, speaker d-vectors are taken as additional inputs to perform speaker adaptive training. The speaker verification network included five convolutional layers. The utterances belonging to the same speaker were concatenated and split into audio segments, each of which had 500 frames. 64-dimensional log Mel-scale filter-bank features were taken as the input.

The d-vector-based SD models were evaluated at both the speaker and utterance levels. During the testing steps, the utterance-level d-vectors were extracted from each utterance separately and the speaker-level d-vectors were extracted using all the utterances from the same speaker. Table 1 reports the word error rate (WER) of the baseline models on SWB task. The performance at the utterance level is much worse than that at the speaker level.

Table 1: *Performance of the baseline models on the SWB task.*

Method	**WER**	**WERR**
SI baseline	13.8	–
SD baseline(speaker-level)	13.0	5.8%
SD baseline(utterance-level)	13.5	2.2%

4.3. Results of the proposed method

For speaker adaptive training method based on the d-vector memory, all speaker-level vectors in the training set were clustered into 128 classes via the K-means algorithm.

Table 2 reports the performance of the proposed method on the SWB task. For our previous work, speaker adaptive training method based on the traditional frame-level attention mecha-

nism achieves a relative 4.3% WER reduction (WERR) compared with the SI model and a relative 2.3% WER reduction over the utterance-level d-vector-based SD model. When we substitute the traditional frame-level attention mechanism with the utterance-level attention-over-attention (AOA) mechanism, the proposed method achieves a relative 8.0% WER reduction over the SI baseline model and a relative 5.9% WER reduction over the utterance-level d-vector-based SD model. In addition, the result of SAT based on AOA mechanism also has lower WER than the speaker-level d-vector-based SD model. If we just average the memory-level attention at all the frames to form the final utterance-level attention directly, no significant improvement can be achieved. For a deep convolution neural network, computational complexity hardly changes at all. Compared with the traditional frame-level attention or utterance-level average attention mechanism, attention-over-attention further strengthen the discrimination among each frame-level speaker embeddings based on the utterance-level long information. In light of this, the generation of the aggregated speaker vector is more robust.

Table 2: *Performance of the proposed method on the SWB task.*

Method	**WER**	**WERR**
SI baseline	13.8	–
SAT with traditional frame-level Att. [18]	13.2	4.3%
SAT with utterance-level average Att.	13.1	5.1%
SAT with utterance-level AOA	12.7	8.0%

We also presented the results on AISHELL-2 task. The results shown in Table 3 are consistent with the results on SWB task. The proposed method achieves a relative 8.3% WER reduction over the SI model and a relative 7.0% WER reduction over the utterance-level d-vector-based SD model.

Table 3: *Performance of the proposed method on the AISHELL-2 task.*

Method	**WER**	**WERR**
SI baseline	7.2	–
SD baseline(speaker-level)	6.9	4.2%
SD baseline(utterance-level)	7.1	1.4%
SAT with traditional frame-level Att. [18]	6.9	4.2%
SAT with utterance-level AOA	6.6	8.3%

To verify the improvement of the speaker embeddings, we compared the aggregated speaker d-vectors with the utterance-level d-vectors with t-distributed stochastic neighbor embedding (t-SNE) [29] on test data. 10 utterances from the same speaker were first randomly picked, and then the utterance-level d-vectors based on attention-over-attention were obtained directly during the attention module. And the utterance-level d-vectors based on the traditional frame-level attention could be generated by averaging all the frame-level aggregated speaker vectors in an utterance. For comparison, traditional utterance-level d-vectors were also extracted for each utterance of each speaker.

As shown in Fig. 3, the aggregated speaker vectors based on different attention mechanism are all closer to the speaker-level d-vector than the traditional utterance-level ones. And compared with the traditional frame-level attention mechanism, the

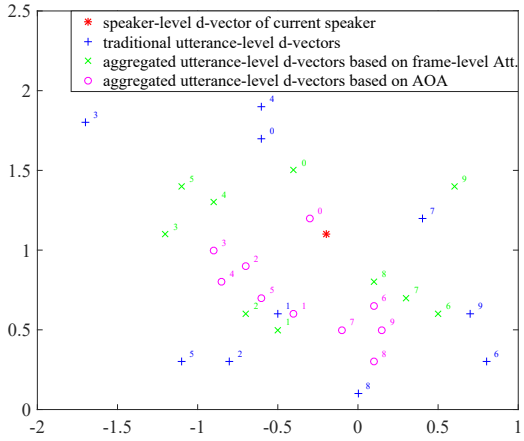

Figure 3: *t-SNE of the different speaker vectors in test set.*

aggregated speaker vectors based on attention-over-attention are more concentrated, which indicates the offsets influenced by the frame with little speaker information are effectively reduced. In order to solidify the conclusion, we calculated the euclidean distance between the utterance-level speaker vector and speaker-level d-vector for all the utterances of all the speakers in the test set, and got the mean and variance of the distance. As shown in Table 4, compared with the aggregated utterance-level d-vectors based on the traditional frame-level attention mechanism, the mean of euclidean distance of the aggregated utterance-level d-vectors based on attention-over-attention mechanism is relatively closer while the variance of the euclidean distance has great advantages. The results prove the superiority of the attention-over-attention mechanism again.

Table 4: *The mean and variance of euclidean distance between different utterance-level d-vectors and speaker-level d-vectors.*

Utterance-level d-vectors	**Mean**	**Variance**
traditional d-vectors	1.43	0.13
aggregated d-vectors based on Att.	1.29	0.10
aggregated d-vectors based on AOA	1.24	0.06

5. Conclusions

In this study, we have proposed a speaker adaptive training method for speech recognition with attention-over-attention mechanism, which can be used to measure the speaker information contribution of each memory vector and each frame. Thus, the utterance-level aggregated vectors are more representative and stable. The results on Switchboard and AISHELL-2 task show that our proposed approach can achieve relative word error rate reductions of 8.0% and 8.3% compared with the speaker independent model respectively, and 6.0%-7.0% compared to that of the traditional utterance-level d-vector-based SAT method. The utterance-level aggregated speaker vectors based on attention-over-attention mechanism yielded relative word error rate reductions of approximately 4.0% compared with the frame-level attention mechanism.

6. References

[1] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of Interspeech*, 2014, pp. 338–342.

[2] O. Abdel-Hamid, A. rahman Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Proceedings of ICASSP*, 2012, pp. 4277–4280.

[3] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. Lecun, "Very deep multilingual convolutional neural networks for lvcsr," in *Proceedings of ICASSP*, 2016, pp. 4955–4959.

[4] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proceedings of Interspeech*, 2014, pp. 2180–2184.

[5] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proceedings of ASRU*, 2013, pp. 55–59.

[6] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 1938–1949, 2015.

[7] P. Cardinal, N. Dehak, Y. Zhang, and J. Glass, "Speaker adaptation using the i-vector technique for bottleneck features," in *Proceedings of Interspeech*, 2015, pp. 2867–2871.

[8] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code," in *Proceedings of ICASSP*, 2013, pp. 7942–7946.

[9] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code," in *Proceedings of ICASSP*, 2014, pp. 6389–6393.

[10] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1713–1725, 2014.

[11] X. Cui, V. Goel, and G. Saon, "Embedding-based speaker adaptive training of deep neural networks," in *Proceedings of Interspeech*, 2017, pp. 122–126.

[12] Y. Zhao, J. Li, S. Zhang, L. Chen, and Y. Gong, "Domain and speaker adaptation for cortana speech recognition," in *Proceedings of ICASSP*, 2018, pp. 5984–5988.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *JMLR Workshop and Conference Proceedings*, vol. 37, pp. 1180–1189, 2015.

[14] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proceedings of NIPS*, 2016, pp. 343–351.

[15] Z. Meng, J. Li, Y. Gong, and B. Juang, "Adversarial teacher-student learning for unsupervised domain adaptation," in *Proceedings of ICASSP*, 2018, pp. 5949–5953.

[16] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gong, and B. Juang, "Speaker-invariant training via adversarial learning," in *Proceedings of ICASSP*, 2018, pp. 5969–5973.

[17] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, and et al., "English conversational telephone speech recognition by humans and machines," in *Proceedings of Interspeech*, 2017.

[18] J. Pan, D. Liu, G. Wan, J. Du, Q. Liu, and Z. Ye, "Online speaker adaptation for lvcsr based on attention mechanism," in *Proceedings of APSIPA*, 2018, pp. 183–186.

[19] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[20] E. Variani, X. Lei, E. McDermott, I. Lopez, Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of ICASSP*, 2014, pp. 4052–4056.

[21] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *Computer Science*, 2015.

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Computer Science*, pp. 2048–2057, 2015.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N, Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *In Advances in Neural Information Processing Systems*, pp. 6000–6010, 2017.

[24] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," *ACL 2017*, pp. 593–602, 2017.

[25] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[26] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," in *CoRR*, 2018.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *CoRR*, 2014.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, 2014, pp. 675–678.

[29] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.