SEQ2SEQ ATTENTIONAL SIAMESE NEURAL NETWORKS FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Yichi Zhang[1*], *Meng Yu*[2], *Na Li*[2], *Chengzhu Yu*[3], *Jia Cui*[3], *Dong Yu*[3]

[1]Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627, USA
[2]Tencent AI Lab, Shenzhen, China [3]Tencent AI Lab, Bellevue, WA 98004, USA

ABSTRACT

In this paper, we present a Sequence-to-Sequence Attentional Siamese Neural Network (Seq2Seq-ASNN) that leverages temporal alignment information for end-to-end speaker verification. In prior works of speaker discriminative neural networks, utterance-level evaluation/enrollment speaker representations are usually calculated. Our proposed model, utilizing a sequence-to-sequence (Seq2Seq) attention mechanism, maps the frame-level evaluation representation into enrollment feature domain and further generates an utterance-level evaluation-enrollment joint vector for final similarity measure. Feature learning, attention mechanism, and metric learning are jointly optimized using an end-to-end loss function. Experimental results show that our proposed model outperforms various baseline methods, including the traditional i-Vector/PLDA method, multi-enrollment end-to-end speaker verification models, d-vector approaches, and a self attention model, for text-dependent speaker verification on a Tencent internal voice wake-up dataset.

Index Terms— End-to-end speaker verification, text-dependent, Siamese neural networks, Seq2Seq attention

1. INTRODUCTION

Speaker verification is the process of verifying, based on a speaker's enrolled utterances, whether an evaluation utterance belongs to that speaker. It can be categorized into text-dependent and text-independent tasks [1]. In text-dependent systems, transcripts of enrollment are constrained to a specific phrase [2], which is not the case in text-independent systems. Because of the constraint of the phonetic variability, text-dependent speaker verification usually achieves robust verification results with very short enrollment utterances. With the proliferation of smart home/vehicles and mobile applications, human-machine interactions through voice command are becoming widespread where text-dependent speaker verification is essential. For example, an ideal application scenario would be speech assisted devices continuously listening for specific wake-up keywords only by a certain speaker, where text-dependent speaker verification is necessary for personalized service and unauthorized usage prevention.

Traditional techniques for text-dependent speaker verification include GMM-UBM [3], GMM-SVM [4], and i-Vector/PLDA [5]. Recently, inspired by the huge success of applying Deep Neural Networks (DNN) in Automatic Speech Recognition (ASR) [6], deep learning based text-dependent speaker verification has become popular. In [2, 7], speaker discriminative DNNs are investigated to extract frame-level features, which are treated with equal importance and aggregated into reliable utterance-level speaker representations called d-vectors. Utterance-level features from the test speaker and enrolled speakers are then scored using a predefined cosine distance [8] or PLDA [9] similarity measure.

The end-to-end text-dependent speaker verification system has also attracted much attention due to its simple training procedure and effective inference scheme. In [10], the last frame output of LSTM layer is defined as the d-Vector for evaluation and enrollment representations, respectively, which are then passed to calculate cosine distance and logistic regression for the similarity score. In [11], a normalized score of each LSTM frame is calculated and all frames are weighted averaged to generate the d-Vector. Similar attention mechanism is applied to a triplet loss model in [12]. Another attention based model in [13] takes the additional phonetic model information to learn the attention weights for each evaluation and enrollment. However, in [11, 12, 13], evaluation and enrollment implement their own attention mechanism and no evaluation-enrollment joint information is utilized.

For a better end-to-end training, the mismatch in the phonetic contexts and duration between the evaluation and enrollment can be resolved by a sequence-to-sequence (Seq2Seq) temporal alignment. Original Seq2Seq attention is widely used in machine translation [14] and image captioning [15], where alignments are learned between source and target sequences. This motivates us to learn temporal alignment between enrollment and evaluation utterances.

In this paper, we propose a Seq2Seq style attentional Siamese neural network model, named Seq2Seq-ASNN, for the above purpose. A Siamese neural network consists of two towers with identical structures for encoding individual

*This work was done while Y. Zhang was an intern at Tencent AI Lab, Bellevue, WA 98004, USA.

input features. It has been successfully applied to many image/video/audio tasks such as face verification [16], object tracking in videos [17], and sound search by vocal imitation [18, 19]. The proposed Siamese neural network encodes an enrollment and an evaluation utterance with separate towers. Each tower is composed of a convolutional layer followed by a recurrent layer to extract the temporal-frequency feature representation. Then the extracted frame-level features from the two towers are weighted aligned and combined into an utterance-level evaluation-enrollment joint vector by an Seq2Seq attention mechanism. The dual-tower feature extraction, Seq2Seq attention mechanism, and the verification scoring are jointly trained by optimizing the end-to-end loss.

The rest of the paper is organized as follows: We describe the proposed Seq2Seq-ASNN in Section 2. The experimental setup is summarized in Section 3. We present the evaluation results in Section 4 and conclude this paper in Section 5.

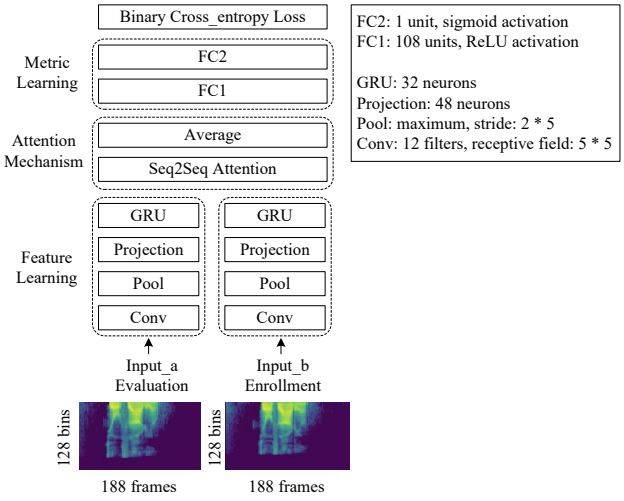

Fig. 1. Architecture of the proposed Seq2Seq-ASNN model for end-to-end speaker verification.

2. THE PROPOSED SEQ2SEQ-ASNN

The typical speaker verification protocol includes three phases: training, enrollment, and evaluation [10]. In the training phase, our proposed network learns to extract internal speaker representations from a pair of utterances. The encoding network includes two parts, a CRNN (CNN + GRU) and an attention network as shown in Figure 1. After feature extraction by the CRNN, the Seq2Seq attention mechanism takes frame-level features to compute attention weights for temporal alignment between evaluation and enrollment representations. Finally, two fully connected layers produce a binary decision on whether the two utterances belong to the same speaker. All the parameters in the whole system are jointly trained using an end-to-end criterion on positive (i.e., two input utterances share the same speaker identity, a.k.a. target samples in testing phase) and negative (i.e., two input utterances belong to different speakers, a.k.a. impostor samples in testing phase) pairs, as described in Section 2.5.

While the attention model in [13] is learned based on individual utterance, our attention model is trained in a Seq2Seq manner where both evaluation and enrollment frame-level features are required to produce an utterance-level joint vector. Besides, although the enrollment and verification phases are implemented in one-shot, enrollment frame-level features could still be extracted and saved beforehand for real-time verification deployment. Finally, in end-to-end settings like [10] and [13], the evaluation and enrollment branches are combined at the "Metric Learning" stage, while our proposed method couples the two branches at the "Attention Mechanism" stage to generate the utterance-level joint vector.

2.1. Preprocessing

The evaluation and enrollment utterances are sampled at 16 kHz and recorded for shorter than 3 seconds. Each utterance is zero-padded in the end to maintain 3 seconds long, and then converted to a 128-band log-mel spectrogram with 32 ms analysis window and 16 ms overlap, resulting in a dimensionality of 128 frequency bins by 188 time frames.

2.2. Feature Learning

Each tower of the Siamese network comprises of a convolutional layer and a recurrent layer. The model parameters are shown on the upper right side of Figure 1. The convolutional layer has 12 filters with Rectified Linear Unit (ReLU) activations and a receptive field of 5×5, followed by a $2(frequency) \times 5(time)$ max-pooling. For each time step, we concatenate the features across different channels, then project to a 48 dimensional layer, and finally feed into a GRU layer with 32 hidden units. Up to now, we get the frame-level features for both evaluation and enrollment utterances.

2.3. Seq2Seq Attention Mechanism

Rather than averaging the frame-level CRNN features to produce an utterance-level representation for enrollment and evaluation respectively, we adopt Seq2seq attention mechanism to first align these two frame-level feature sequences. Particularly, each enrollment frame can be aligned to a weighted average of evaluation frames. This average representation is then concatenated with the original enrollment feature to form a unified feature sequence of the two utterances, which is then averaged to generate an evaluation-enrollment joint vector.

The internal structure for the proposed Seq2Seq attention mechanism is shown in Figure 2. $\mathbf{h_s}$ and $\mathbf{h_t}$ are the evaluation and enrollment frame-level speaker feature sequences, respectively. Our goal is to derive a context vector sequence

$\mathbf{c_t}$ that captures evaluation side information that is the most enrollment-relevant. By defining an alignment score between the s-th frame of the evaluation and the t-th frame of the enrollment as the following:

$$score(\mathbf{h_t}, \mathbf{h_s}) = \mathbf{h_t}^\top \mathbf{h_s}, \tag{1}$$

we can find the frame-level relevance between the two sequences. The higher alignment score indicates a larger weight of that evaluation frame contributing to the context vector. The variable-length alignment vector $\alpha_\mathbf{t}$ can be realized by a softmax function as:

$$\alpha_\mathbf{t}(\mathbf{s}) = \frac{\mathbf{exp}(\mathbf{score}(\mathbf{h_t}, \mathbf{h_s}))}{\sum_{\mathbf{s'}}^{\mathbf{S}} \mathbf{exp}(\mathbf{score}(\mathbf{h_t}, \mathbf{h_{s'}}))}, \tag{2}$$

where the size of $\alpha_\mathbf{t}(\mathbf{s})$ equals the number of time steps on the evaluation side. Given the alignment vector as weights, the context vector $\mathbf{c_t}$ is computed as the weighted average over all the evaluation frame-level speaker vectors:

$$\mathbf{c_t} = \sum_{\mathbf{s}}^{\mathbf{S}} \alpha_\mathbf{t}(\mathbf{s}) \mathbf{h_s}. \tag{3}$$

Hence, the context vector sequence $\mathbf{c_t}$ has the same number of time steps with the enrollment sequence $\mathbf{h_t}$. We employ a concatenation layer to combine the information from both vectors to produce an attentional hidden state $\tilde{\mathbf{h}}_t$. As such, the frame vectors in the evaluation are automatically weighted aligned to the highly correlated frames from the enrollment. Finally, we average the vectors $\tilde{\mathbf{h}}_t$ across all time steps to get a 32-d utterance-level joint vector as the output of the Seq2Seq attention module, which contains the integrated information from both evaluation and enrollment utterances.

2.4. Metric Learning

Instead of calculating pre-defined distances between the evaluation/enrollment utterance-level feature representations in [10], we feed the joint vector learned from the Seq2Seq attention mechanism through a 2-layer Fully Connected Network (FCN) to predict the evaluation-enrollment pair similarity. The FC1 layer consists of 108 hidden units using rectified linear unit (ReLU) nonlinearity, followed by a sigmoid output layer of only one neuron for verification score prediction. The similarity prediction is learned jointly with the feature representations and attention mechanism, likely leading to a better speaker verification performance.

2.5. Training

For the end-to-end training scheme, the ground truth labels are 1 for positive pairs and 0 for negative pairs. The loss function to minimize is the binary cross-entropy between the network output and the binary label. The learning rate of stochastic gradient descent optimization is 0.1 with a decay rate 0.001 and a momentum constant of 0.9. The batch size is 256. Early stopping based on validation loss with the patience of 3 epochs is employed for training termination.

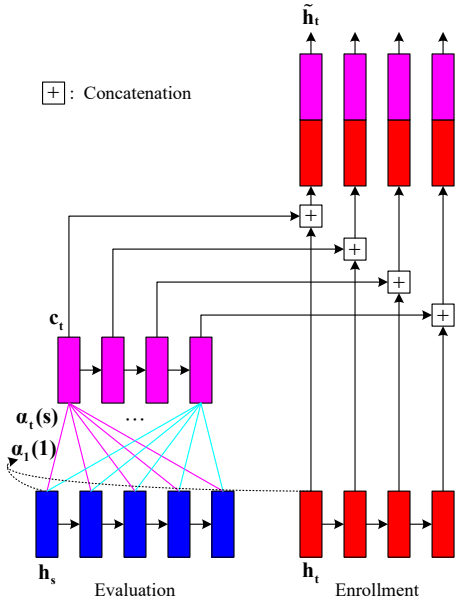

Fig. 2. Structure of the Seq2Seq attention mechanism. Each enrollment frame is aligned to a weighted average of evaluation frames.

3. EXPERIMENTAL SETUP

3.1. Dataset

We use a Tencent wake-up word dataset in our experiments. It contains 3,324 speakers of equal gender representation. 30 utterances of the keyword "9420" spoken in Chinese are recorded for each speaker. We split the entire dataset into 2,570, 635, and 119 speakers for training, validation, and testing, respectively. By pairing utterances from the same speaker as positive samples and from different speakers as negative samples, a total number of 74k positive samples and 74k negative samples are created for training. This number is about 18k for validation. In the testing phase, there are in total 12k target and impostor samples.

3.2. Baseline Methods

We compare the proposed method with several baseline speaker verification approaches.

(1) Traditional i-Vector/PLDA method [5] that adopts 512-d i-vectors and reduces to 200-d by LDA. And then a PLDA model with 150 latent identity factors is trained.

(2) Google's end-to-end text-dependent speaker verification system (Google-E2E) [10] that receives an evaluation and multiple enrollments (here we choose three) as inputs. It first extracts features from the evaluation and enrollment utterances by a LSTM layer with 504 hidden neurons, then calculates the cosine similarity between the evaluation representation and the averaged enrollment representation, which

is finally fed into a logistic regression to generate a similarity score. For the input log-mel spectrogram dimensionality, we perform experiments with both Google-E2E-1 with $40(frequency) \times 80(time)$ as described in [10] and Google-E2E-2 with $128(frequency) \times 188(time)$, which has the same input dimensionality as Seq2Seq-ASNN.

(3) The d-Vector approach [2]. An utterance is classified to one of 2,570 training speakers with a softmax output. The last hidden layer activation is used as the evaluation and enrollment d-Vector, respectively. Then cosine distance between the two parties is calculated. Distances lower than the threshold suggest positive pairs, otherwise negative. Another improvement further applies the Cosface normalization to the last hidden layer together with a large margin loss [20].

(4) Self attention. Unlike [11], we apply a more complicated self attention described in [21] to replace the Seq2Seq attention, named Self-ASNN. It implements individual attention for evaluation and enrollment separately. Weights are calculated as activations of a fully connected layer following GRU output across frames. Re-weighted frames are obtained by multiplying weights with GRU outputs, which are further averaged into an utterance-level feature vector for each tower. Finally evaluation and enrollment feature vectors are concatenated and fed into the FCN for similarity measure.

(5) Comparison of Seq2Seq-ASNN against itself by gradually removing attention and GRU layers. Siamese-CNN-GRU removes the attention layer, concatenates the last frame GRU output from evaluation and enrollment, which is then fed into FCN for similarity measure. Siamese-CNN removes both GRU and attention layers but employs three convolutional layers, maintaining a comparable model complexity with Seq2Seq-ASNN and Siamese-CNN-GRU.

4. EXPERIMENTAL RESULTS

We employ Equal Error Rate (EER) to evaluate the speaker verification performance. We also report the model size in terms of the number of trainable parameters for deep learning models. Table 1 shows EER and model size comparisons among the proposed Seq2Seq-ASNN and baseline methods.

The traditional i-Vector/PLDA method achieves the best result among all baseline approaches. Unlike data driven deep learning based models, i-Vector/PLDA is robust under the circumstance of limited number of training speakers. However, Seq2Seq-ASNN outperforms the best baseline by a relative EER decrease of 35.7%. This indicates the effectiveness of the proposed Siamese neural network structure as well as the Seq2Seq attention mechanism. It may suggest that if more training data is available, a more significant performance improvement of Seq2Seq-ASNN could be achieved.

With a smaller model size, Seq2Seq-ASNN outperforms both Google-End2End and d-Vector approaches. For Google-End2End models, we find the benefit from a larger input spectrogram resolution. However, even with three enrollments likely for a better temporal feature coverage, Google-End2End still performs worse than Seq2Seq-ASNN, which requires one enrollment. This may suggest that in Google-End2End models averaged enrollment representation makes the temporal mapping with the evaluation utterance even harder to capture. Also it confirms the effectiveness of the Seq2Seq attention mechanism. The d-Vector method, gaining huge from the Cosface loss, is still outpaced by Seq2Seq-ASNN. This indicates the performance of d-Vector method is limited by the relatively smaller training set, while Seq2Seq-ASNN learns pairwise relative speaker features which are not directly related to absolute speaker identities.

Table 1. EER and model size (# trainable parameters) comparisons of Seq2Seq-ASNN with various baseline systems.

Configuration	Model Size	EER
i-Vector/PLDA	-	0.56%
Google-End2End-1 (40×80)	1.1M	4.56%
Google-End2End-2 (128×188)	1.1M	4.28%
d-Vector (w/o Cosface)	0.3M	8.00%
d-Vector (w/ Cosface)	0.3M	1.50%
Self-ASNN	149.7k	1.73%
Siamese-CNN	146.7k	3.40%
Siamese-CNN-GRU	148.4k	1.87%
Seq2Seq-ASNN	149.7k	**0.36%**

Self-ASNN performs better than Siamese-CNN-GRU. It suggests that frame re-weighting within evaluation and enrollment does have benefit. However, Seq2Seq attention still outperforms self attention with a large margin. This indicates that temporal alignment between enrollment and evaluation utterance pair is essential for achieving improved speaker verification performance.

By better capturing local time-frequency features as well as long timescale temporal evolutions, Siamese-CNN-GRU outperforms Siamese-CNN. Furthermore, Seq2Seq-ASNN outperforms Siamese-CNN-GRU by a decrease of EER from 1.87% to 0.36%. This suggests that Seq2Seq attention works as expected for evaluation/enrollment temporal alignment.

5. CONCLUSIONS

In this paper, we propose a Seq2Seq-ASNN model for text-dependent speaker verification. Seq2Seq attention maps frame-level evaluation features into enrollment domain and generates an utterance-level evaluation-enrollment joint vector for similarity measure. Feature extraction, attention mechanism, and metric learning are jointly optimized in an end-to-end manner. Experimental results show significant improvement of Seq2Seq-ASNN against various baselines on a Tencent wake-up word dataset. For future work, we will evaluate our system on publicly available corpus and apply the proposed method to text-independent speaker verification.

6. REFERENCES

[1] Joseph P Campbell, "Speaker recognition: A tutorial," *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. Acoust., Speech and Signal Process. (ICASSP), 2014 IEEE Int. Conf. on*, 2014, pp. 4052–4056.

[3] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal process.*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] William M. Campbell, Douglas E. Sturim, and Douglas A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, 2006.

[5] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[7] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: An end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[8] Najim Dehak, Reda Dehak, James R Glass, Douglas A Reynolds, and Patrick Kenny, "Cosine similarity scoring without score normalization techniques," in *Odyssey*, 2010, pp. 1–5.

[9] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md Jahangir Alam, and Pierre Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. Acoust., Speech and Signal Process. (ICASSP), 2013 IEEE Int. Conf. on*, 2013, pp. 7649–7653.

[10] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Proc. Acoust., Speech and Signal Process. (ICASSP), 2016 IEEE Int. Conf. on*, 2016, pp. 5115–5119.

[11] F. A. Chowdhury, Quan Wang, Ignacio Lopez Mereno, and Li Wan, "Attention-based models for text-dependent speaker verification," *arXiv preprint arXiv:1710.10470*, 2017.

[12] Subhadeep Dey, Srikanth Madikeri, and Petr Motlicek, "End-to-end text-dependent speaker verification using novel distance measures," in *Proc. Interspeech*, 2018, pp. 3598–3602.

[13] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong, "End-to-end attention based text-dependent speaker verification," in *Proc. Spoken Language Technol. (SLT)*, 2016, pp. 171–178.

[14] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proc. Comput. Vision and Pattern Recognition (CVPR), 2015 IEEE Conf. on*, 2015, pp. 3156–3164.

[16] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. Comput. Vision and Pattern Recognition (CVPR)*, 2005, pp. 539–546.

[17] Luca Bertinetto, Jack Valmadre, Joao F. Henriques, Andrea Vedaldi, and Philip HS Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. on Comput. Vision (ECCV)*, 2016, pp. 850–865.

[18] Yichi Zhang and Zhiyao Duan, "IMINET: Convolutional semi-Siamese networks for sound search by vocal imitation," in *Proc. Appl. of Signal Process. to Audio and Acoust. (WASPAA), 2017 IEEE Workshop on*, 2017, pp. 304–308.

[19] Yichi Zhang and Zhiyao Duan, "Visualization and interpretation of Siamese style convolutional neural networks for sound search by vocal imitation," in *Proc. Acoust., Speech and Signal Process. (ICASSP), 2018 IEEE Int. Conf. on*, 2018, pp. 2406–2410.

[20] Hao. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," *arXiv preprint arXiv:1801.09414*, 2018.

[21] Philippe Rémy, "Keras Attention Mechanism," `https://github.com/philipperemy/keras-attention-mechanism/`, 2017, [Online; accessed October 27, 2018].