

Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings

Jee-weon Jung*, Hee-Soo Heo*, Hye-jin Shim, and Ha-Jin Yu†

School of Computer Science, University of Seoul, South Korea

jeewon.leo.jung@gmail.com, zhasgone@naver.com, shimhz6.6@gmail.com, hjyu@uos.ac.kr

Abstract

The short duration of an input utterance is one of the most critical threats that degrade the performance of speaker verification systems. This study aimed to develop an integrated text-independent speaker verification system that inputs utterances with short duration of 2 seconds or less. We propose an approach using a teacher-student learning framework for this goal, applied to short utterance compensation for the first time in our knowledge. The core concept of the proposed system is to conduct the compensation throughout the network that extracts the speaker embedding, mainly in phonetic-level, rather than compensating via a separate system after extracting the speaker embedding. In the proposed architecture, phonetic-level features where each feature represents a segment of 130 ms are extracted using convolutional layers. A layer of gated recurrent units extracts an utterance-level feature using phonetic-level features. The proposed approach also adopts a new objective function for teacher-student learning that considers both Kullback-Leibler divergence of output layers and cosine distance of speaker embeddings layers. Experiments were conducted using deep neural networks that take raw waveforms as input, and output speaker embeddings on VoxCeleb1 dataset. The proposed model could compensate approximately 65 % of the performance degradation due to the shortened duration.

Index Terms: Short utterance compensation, teacher-student learning, text-independent speaker verification, raw waveform, speaker embedding

1. Introduction

Recent speaker verification systems generally work based on utterance-level features such as i-vectors, or speaker embeddings from deep neural networks (DNNs) such as x-vector system [1–3]. In utterance-level features extracted from short utterances, uncertainty exist owing to the insufficient phonetic information, which is a well-known factor of performance degradation of speaker verification systems [4]. To compensate for this uncertainty caused by short utterances, Saeidi *et al.* proposed a propagation method in the i-vector space [5]. Yamamoto *et al.*, proposed a DNN-based compensation system that transforms an i-vector extracted from a short utterance into an i-vector corresponding to a long utterance. In Yamamoto’s research, it was shown that phonetic information can be effectively used for compensating short utterances [6]. However, only a minor improvement in performance could be obtained through this approach. We assume that this limitation occurred because it is

difficult to compensate the missing phonetic information using already extracted utterance-level features [5–7].

Unlike most previous studies that compensate utterance-level features after they have been extracted, we propose a novel short utterance compensation system based on phonetic-level features. The proposed system extracts speaker embeddings directly from short utterances. The phonetic-level feature in this study is defined as an intermediate concept between frame-level and utterance-level features that effectively represents phonetic information, which covers approximately 130 ms. The duration of 130 ms is known to be appropriate for representing phonetic information based on conventional phonetic knowledge [8, 9]. Figure 1-(a) illustrates the concept of the phonetic-level features.

To efficiently compensate the short utterances using phonetic information, we use the convolutional neural network long short-term memory (CNN-LSTM) architecture proposed by Jung *et al.* with a few modifications [10]. This model directly extracts utterance-level embeddings from raw waveform, where the process can be divided into frame-level feature extraction and utterance-level feature aggregation. The CNN is used to conduct the former and LSTM is used for the latter. Here, we define the output of the last convolutional layer as the phonetic-feature. Using this model, teacher-student (TS) learning framework is conducted where cosine distance of speaker embeddings from long and short utterances are compared to efficiently compensate the short utterances using phonetic information. Resulting proposed system is an integrated short utterance compensation system that extracts speaker embeddings directly from short utterances of 2.05 s duration, text-independently (overall illustration in Figure 1-(b)).

The remaining paper is organized as follows: Section 2 describes the speaker embedding system. Section 3 introduces the teacher-student learning framework. The proposed short utterance compensation system is discussed in Section 4. The experimental settings and result analysis are described in Section 5 and the study is concluded in Section 6.

2. Speaker embedding model

Recent advances in deep neural networks (DNNs) have resulted in several successful speaker embedding systems that directly model raw waveforms [10–12]. These studies have shown that suitable pre-processing for speaker verification could be performed, yielding comparable or better results than conventional Mel-energy feature or spectrogram-based systems [10, 13]. In this study, we use the raw waveform CNN-LSTM (RWCNN-LSTM) architecture proposed in [10] with the following two modifications: leaky rectified linear unit (LReLU) activation was used [14] instead of ReLU activation, and the long short-term memory layer was replaced with a GRU layer. Comparative experimental results show that these two modifications lead

*These authors contributed equally

† Corresponding author

This work was supported by the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea)

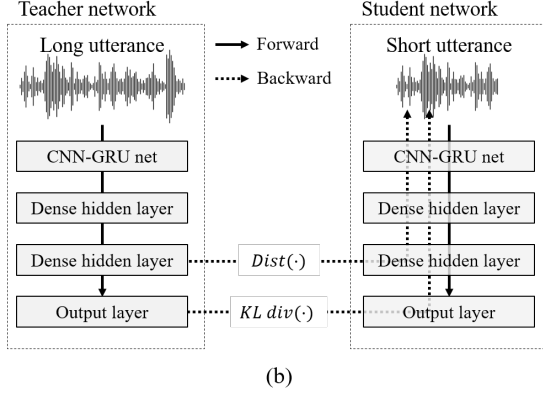
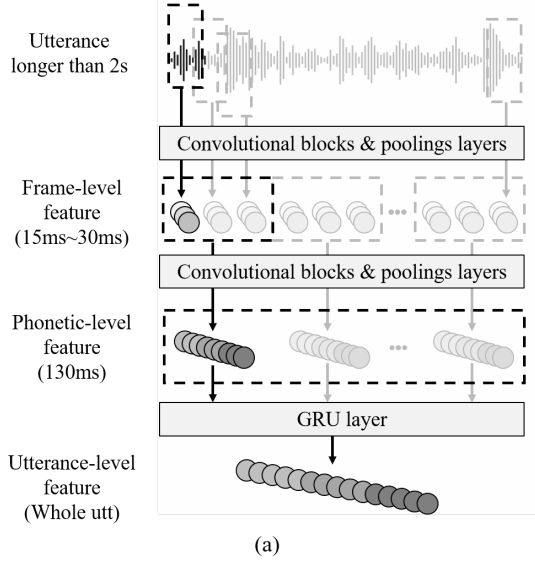


Figure 1: (a) Conceptual illustration of various levels of features based on CNN-GRU network (b) Workflow of the proposed teacher-student learning-based short utterance compensation system.

to an additional decrease of 10 % in terms of equal error rates (EERs). Note that other DNN systems can also be used for the short utterance compensation scheme proposed in Section 4.

The RWCNN-GRU model comprises convolutional blocks followed by one GRU layer and two fully-connected layers. Each convolutional block has a residual connection [15] and contains a pooling layer at its end, same with the convolutional block used in [10]. The last convolutional block transforms raw waveforms into phonetic-level features (addressed in Section 4) that represent segments of approximately 130 ms. The outputs of the last convolutional block are fed to the following GRU layer, which produces fixed low dimensional utterance-level representations. Then, the last fully-connected layer’s LReLU activation is used as the speaker embedding. Speaker verification is performed by comparing the cosine distance between two speaker embeddings. In this research, both teacher and student DNNs have an identical architecture. However, the sequence length of the output of the last convolutional block (which can also be thought of the timestep of the GRU input) varies depending on the length of input utterances.



Figure 2: Speaker embeddings visualized using the t-SNE algorithm [19]. The five different colors represent five randomly selected speakers from the evaluation set. A triangle denotes the mean of the speaker embeddings extracted from long utterances.

3. Teacher-student learning

Teacher-student (TS) learning uses two DNNs, teacher and student, in which the student DNN is trained using soft labels that the pre-trained teacher DNN provides. In this framework, after the teacher network is trained, the student network is trained to have an output distribution similar to that of the teacher network. This framework was first proposed for model compression and is also widely used for compensating far-field utterances [16–18]. Adoption of the TS framework into short utterance compensation is novel in our knowledge. When TS learning is used for short utterance compensation, the Kullback–Leibler (KL) divergence objective function can be written as

$$KL_{loss} = - \sum_j^J \sum_i^I p_T(s_i|x_{j,l}) \log(p_S(s_i|x_{j,sh})), \quad (1)$$

where i and j refer to the speaker and utterance indices, respectively; x_l and x_{sh} refer to the long and short crop of the same utterance, respectively; and $p_T(s_i|\cdot)$ and $p_S(s_i|\cdot)$ are the outputs of the teacher and student DNNs, respectively. The above equation shows that TS learning trains the student DNN’s output layer distribution same as that of the teacher DNN despite being provided with short utterances.

4. Proposed short utterance compensation system

The main goal of the proposed system is to conduct the compensation throughout the network, rather than compensating via a separate system after extracting the speaker embedding. TS learning addressed in Section 3 is applied to short utterance compensation for this goal. To compensate more efficiently, we propose a direct compensation using the speaker embeddings and the output layer while conventional TS learning only uses the output layer. This is because the ultimate goal of a short utterance compensation system is to make the speaker embedding of a short utterance identical to that of a long utterance,

comparison of speaker embeddings would be a more direct approach. The objective function of the proposed TS learning can be written as an extension of Equation 1,

$$\begin{aligned} Loss = & \sum_j^J Dist(p_T(e|x_{j,t}), p_S(e|x_{j,sh})) \\ & - \sum_j^J \sum_i^I p_T(s_i|x_{j,t}) \log(p_S(s_i|x_{j,sh})). \end{aligned} \quad (2)$$

Here $p_T(e|x_{j,t})$ and $p_S(e|x_{j,sh})$ denote the speaker embedding of the teacher and student DNNs, respectively, and $Dist(\cdot, \cdot)$ denotes the measure of the distance between two embeddings such as the cosine distance or mean squared error.

The approach presented herein is notably different from existing short utterance compensation approaches owing to two aspects. The first is that short utterances are compensated throughout the entire DNN, mainly phonetic-level and GRU layer, rather than being compensated after extracting utterance-level features. Previous researches exploited an additional compensation system to transform speaker embeddings extracted from short utterances after utterance-level feature extraction. This is because the uncertainty caused by lacking phonetic information is observed in utterance-level features. However, compensating phonetic-level features appears to be a more direct solution, because uncertainty arises in the process of extracting utterance-level features from phonetic-level features. We argue that using the proposed approach, although the transformation is performed throughout the network, compensation is mainly conducted in the GRU layer where it tries to move the utterance-level features of the short utterances to the optimal position derived from the corresponding long utterance with abundant phonetic information. Plots in Figure 2 is used to reinforce this argument.

Figure 2 demonstrates the speaker embeddings before (left column), and after (right column) the GRU layer of the baseline (w/o TS, upper row) and the proposed model (w TS, lower row). Because the embeddings are from the evaluation set, unseen data, we expect that cohesiveness of each speaker’s embeddings directly demonstrates the discriminative power. By comparing (a), and (b), we can conclude that the GRU layer increases the discriminative power for each speaker in both baseline and the proposed system. However, Figure 2 shows that the compensation is mainly conducted in the GRU layer because the difference of cohesiveness between (b), (d) is more noticeable than that between (a), (c).

The second difference pertains to the adoption of the approaches of compensating short utterances and maintaining the discriminative power simultaneously using the proposed TS learning approach with the proposed objective function. In [20], Jiachen *et al.* reported that when short utterance compensation is performed, the speaker embedding of the short utterance become close to that of a long utterance. However, even though the distance between short and long utterances became closer in terms of a distance measure, the discriminative power of the compensated embedding could not be ensured. Our experimental results also confirmed that solely reducing the distance between two embeddings of long and short utterances did improve the performance, although not considerably. Therefore, to maintain the discriminative power of the speaker embedding when short utterance compensation is performed, KL-divergence term is included in the proposed objective function. This results in the speaker embedding layers being compared using the cosine distance metric (compensation), while

also using the conventional KL-divergence loss (discriminative power), which is novel. Superior results were obtained using both losses as the final objective function. Overall illustration of the proposed system is depicted in Figure 1-(b).

5. Experiments

5.1. Dataset

In all the experiments described herein, we used the VoxCeleb1 dataset, which comprises approximately 330 hours of audio of 1,251 speakers, at a sampling rate of 16 kHz [21]. The dataset involves utterances with an average and minimum duration of 8.2 s and 4 s, respectively, in a text-independent scenario. Our evaluation trials and training / evaluation subset divisions follow the dataset’s guidelines. To evaluate the performance on the long and short utterances, utterances of the evaluation set were cropped into lengths of 3.59 s (59,049 samples) and 2.05 s (32,805 samples), both enrollment and test utterances. We took the center part of each utterance to compose evaluation sets.

5.2. Experiment configurations

The systems were implemented using Keras, which is a python library with a Tensorflow backend [22–24]. We used the RWCNN-LSTM system [10] with two modifications for both teacher and student DNN architectures. The teacher DNN inputs the raw waveform corresponding to 59,049 samples (≈ 3.59 s). It involves one strided convolutional layer with stride size of 3 and six residual convolution blocks that do not reduce the length of the input sequence (the residual block is identical to that employed in [10]). After each residual convolution block, a max pooling layer with stride size of 3 is applied. The output shape of the last convolution block is (27, 512) where 27 is the sequence length and 512 is the number of kernels in the last convolutional layer. 27 is derived from $59,049 / (3 \times 3^6)$ where 59,049 is the number of samples, 3 is for strided convolution, and 3^6 for six max pooling layers. We note that the number of phonetic-level embeddings extracted using CNN is fixed to 27 in training phase for batch construction of utterances of 59,049 samples, but can vary at evaluation phase depending on the duration of each utterance (e.g. each 2,187 samples yield one phonetic-level embedding, utterance of 2.05 s duration will yield 15 phonetic-level embeddings). The GRU layer has 512 units and the two fully-connected layers have 1,024 nodes each. The multi-step training proposed in [10, 25] is used for training the teacher DNN. The weights of the teacher DNN are frozen when the student DNN is trained.

The student DNN is initialized using the weights of the teacher DNN as this process has been proved to ease the training in [26]. The architecture of the student DNN is identical to that of the teacher DNN except that the student DNN inputs raw waveform with 32,805 samples (≈ 2.05 s), which means that the output shape of the last convolution block is (15, 512). When training the student DNN, two mini-batches where one comprises utterances of 59,049 samples and the other comprises 32,805 samples are respectively fed into teacher and student DNN. Then, cosine distance and KL-divergence are calculated using the last hidden layers and output layers of teacher and student DNN.

The stochastic gradient descent with learning rate of 0.001 and momentum of 0.9 was used as the optimizer when training the teacher DNN. The same optimizer with a learning rate of 0.01 was used for training the student DNN.

5.3. Results and analysis

The baseline performances are presented in Table 1. Using VoxCeleb1 evaluation set without duration restriction which comprises approximately 3 s to 7 s duration, EER of 7.51 % was obtained. EER increased by 46 %, relatively, when the duration of the evaluation set was changed from 3.59 s to 2.05 s, which shows performance degradation owing to the short duration (8.72 % to 12.8 %). The research objective in this study is to compensate EER of 12.8 % to 8.72 %.

Experimental result of training with short utterances at the first time, one of the well-known approaches for short utterance compensation, is shown in the lower row of Table 1. Performance did improve, but only minor improvement of 5 % relative reduction in terms of EER was obtained. This result seems to have occurred because the duration of the short utterance considered herein is less than that used in other studies, in a text-independent scenario (configurations of 5 or 10 s are usually used) [6].

Table 1: *Performance of the baseline systems with different durations. “Full-length eval” corresponds to the use of various length utterances without modification. The numbers represent EERs (%).*

System	full-length eval	3.59 s eval	2.05 s eval
RWCNN-GRU (3.59 s train)	7.51	8.72	12.80
RWCNN-GRU (2.05 s train)	-	-	12.08

Table 2 presents the results of the proposed approaches. Conventional TS learning, which uses the output layer’s KL-divergence loss, did not show noticeable improvement. The proposed method that directly compares the speaker embedding layers demonstrated a better performance (using only the ‘dist’ term in Equation 2), with EER 10.98 % and 10.8 % for mean squared error and cosine distance as distance metrics, respectively. The best result could be achieved by using both the KL-divergence of the output layer and cosine distance of speaker embedding layer, which compensated more than 65 % of the performance degradation due to shortened input utterance. We interpret that the reason for additional performance increase by comparing both output and speaker embedding can be found in Jiachen *et al.*’s research [20]. This research suggested that when compensating short utterances, the compensated feature can become similar to that of the long utterance in terms of the distance scale used for compensation (i.e. Euclidean), but this may not lead to increase in its discriminative power. In other words, in our interpretation, it means that usage of distance metric alone cannot consider the manifold structure of the speaker embedding space. Referring to this argument, comparing speaker embeddings can make the embedding of the student DNN equivalent to that of the teacher DNN, and the KL-divergence between output layers can help maintain its discriminative power.

Additionally, Table 3 shows the evaluation using varying duration utterances. It is not realistic to fix the duration of utterance in real world applications, which makes less duration variant systems necessary. To verify how invariant the proposed system is towards varying duration short utterances, experiments with different range of duration have been conducted. Results demonstrate that EER of both baseline system (w/o TS,

Table 2: *Evaluation of various proposed systems using the modified 2.05 s evaluation set. “Embedding” and “Output” refer to layers to compare between the teacher and student networks. Values inside the bracket indicates the metric.*

Systems	EER (%)
Output (KL-Div) (Original TS)	12.46
Embedding (MSE)	10.98
Embedding (Cos Sim)	10.80
Embedding (Cos Sim)+Output (KL-Div)	10.08

upper) and the proposed system (w TS, lower) are not much variant to the duration of input utterance. We note that the performance degradation as range widens is considered as the effect of inclusion of shorter utterances (e.g. duration starts from 1.55 s in “1.55±0.5 s”).

Table 3: *Evaluation with varying utterance duration. Performance degradation as range widens is due to shorter utterances (e.g. duration starts from 1.55 s in “2.05±0.5 s”).*

System	2.05	2.05 ±0.1 s	2.05 ±0.5 s
RWCNN-GRU (w/o TS)	12.80	12.96	13.28
RWCNN-GRU (w TS)	10.08	10.29	10.40

6. Discussion and future work

In this paper, we proposed a text-independent short utterance speaker verification system that works on utterances with durations of 2.05 s or less. The proposed system does not transform the utterance-level feature from the short utterance as in conventional approaches. Rather, it directly extracts the compensated speaker embeddings from short utterances by compensating throughout the network, focusing on phonetic-level compensation. This is because we expected that the main key for compensating short utterances corresponds to the phonetic information, whose absence leads to the uncertainty of the utterance-level features. To process phonetic information, phonetic-level features that represent segments of 130 ms were extracted using CNN, and then transformed to the utterance-level features using a GRU layer. The effectiveness of the defined phonetic-level features was verified by the performance improvement of the speaker verification system using short utterance compensation and an illustration of the cohesiveness of speaker embeddings from the evaluation set. An objective function is also proposed to conduct a more effective compensation by considering distance of long and short utterance and concurrently maintain the discriminative power of speaker embeddings.

In the future, we will analyze the information included in phonetic-level features and construct phonetic-level features using speech recognition systems. Additionally, because the proposed compensation scheme can be applied to any DNN-based speaker embedding extraction schemes, such as Mel-energy feature-based x-vector or other spectrogram-based systems, we plan to apply the proposed scheme into other systems.

7. References

- [1] 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.
- [2] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.
- [3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [4] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.
- [5] R. Saeidi and P. Alku, "Accounting for uncertainty of i-vectors in speaker recognition using uncertainty propagation and modified imputation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [6] H. Yamamoto and T. Koshinaka, "Denoising autoencoder-based speaker feature restoration for utterances of short duration," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [7] I. Yang, H. Heo, S. Yoon, and H. Yu, "Applying compensation techniques on i-vectors extracted from short-test utterances for speaker verification using deep neural network," in *Proc. ICASSP*. IEEE, 2017.
- [8] G. Peterson and I. Lehiste, "Duration of syllable nuclei in english," *The Journal of the Acoustical Society of America*, vol. 32, no. 6, pp. 693–703, 1960.
- [9] M. Ordin and L. Polyanskaya, "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 533–544, 2015.
- [10] J. Jung, H. Heo, I. Yang, H. Shim, and H. Yu, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," in *Proc. Interspeech 2018*, 2018, pp. 3583–3587.
- [11] —, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5349–5353.
- [12] H. Muckenhirn, M. Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using cnns," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4884–4888.
- [13] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with sincnet," *arXiv preprint arXiv:1811.09725*, 2018.
- [14] A. L. Maas, A. Y. Hannun, and A. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [16] J. Li, R. Zhao, J. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth annual conference of the international speech communication association*, 2014.
- [17] J. Li, R. Zhao, Z. Chen, C. Liu, X. Xiao, G. Ye, and Y. Gong, "Developing far-field speaker system via teacher-student learning," *arXiv preprint arXiv:1804.05166*, 2018.
- [18] J. Kim, M. El-Khamy, and J. Lee, "Bridgenets: Student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5719–5723.
- [19] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, 2008.
- [20] J. Zhang, N. Inoue, and K. Shinoda, "I-vector transformation using conditional generative adversarial networks for short utterance speaker verification," *Proceedings of INTERSPEECH, Hyderabad, India*, 2018.
- [21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.
- [22] F. Chollet *et al.*, "Keras," <https://github.com/keras-team/keras>, 2015.
- [23] A. Martín, A. Ashish, B. Paul, B. Eugene *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015. [Online]. Available: <http://download.tensorflow.org/paper/whitepaper2015.pdf>
- [24] A. Martin, B. Paul, C. Jianmin, C. Zhifeng, D. Andy, D. Jeffrey, D. Matthieu, G. Sanjay, I. Geoffrey, I. Michael, K. Manjunath, L. Josh, M. Rajat, M. Sherry, M. G. Derek, S. Benoit, T. Paul, V. Vijay, W. Pete, W. Martin, Y. Yuan, and Z. Xiaoqiang, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>
- [25] H. S. Heo, J. W. Jung, I. H. Yang, S. H. Yoon, and H. J. Yu, "Joint training of expanded end-to-end DNN for text-dependent speaker verification," *Proc. Interspeech 2017*, pp. 1532–1536, 2017.
- [26] R. Pang, T. Sainath, R. Prabhavalkar, S. Gupta, Y. Wu, S. Zhang, and C. Chiu, "Compression of end-to-end models," 2018.