An End-to-End Text-Independent Speaker Identification System on Short Utterances

Ruifang Ji[1,2], *Xinyuan Cai*[1], *Bo Xu*[1]

[1]Institute of Automation, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China

{jiruifang2016, xinyuan.cai, xubo}@ia.ac.cn

Abstract

In the field of speaker recognition, text-independent speaker identification on short utterances is still a challenging task, since it is rather tough to extract a robust and dicriminative speaker feature in short duration condition. This paper explores an end-to-end speaker identification system, which maps utterances to a speaker identity subspace where the similarity of speakers can be measured by Euclidean distance. To be specific, we apply stacked gated recurrent unit (GRU) architectures to extract utterance-level feature. Then it is assumed that one's various utterances can be viewed as transformations of a single object in an ideal speaker identity subspace. Based on this assumption, the residual convolution neural network (ResCNN) architecture is utilized to model the transformation, and the whole system is jointly optimized by speaker identity subspace loss. Experimental results demonstrate the effectiveness of our proposed system and superiroity over pervious methods. For example, the GRU learned feature reduces the equal error rate by 27.53% relatively and the speaker identity subspace loss further brings 7.22% relative reduction compared to softmax loss.

Index Terms: speaker identification, short duration, speaker identity subspace loss, GRU, ResCNN

1. Introduction

Speaker identification is the process of classifying the identity of an unknown voice among a set of speakers, based on the speaker's known utterances. Depending on the restrictions of the utterances, speaker identification models usually fall into two categories: text-dependent speaker identification (TD-SI) and text-independent speaker identification (TI-SI). When the transcript of utterances is lexically constrained, the task is considered as TD-SI, otherwise it is TI-SI.

The traditional speaker recognition approach, i-vectors systems, has been dominant for years [1, 2]. Its framework contains three stages [3]: a model (e.g., UBM, DNN) to collect Baum-Welch statistics, a projection matrix to convert high-dimensional statistics to a single low-dimensional speaker embedding (i-vector), and a classifing backend (PLDA) to produce similarity scores by comparing i-vectors of different utterances. Despite its success, it suffers from a major drawback. That is, the subtasks are trained individually without tight connection.

Using different deep learning frameworks with end-to-end loss functions to train speaker discriminative embeddings, has drawn much attention recently [4, 5, 6]. In [4], DNNs with network-in-network (NIN) [7] model and a PLDA based loss function, achieved better speaker verification performances when 105k speakers were employed in the network training. In [5], long short-term memory (LSTM) [8] model with a logistic regression function based end-to-end loss, reached a 2% equal error rate (EER) on the "Ok, Google" benchmark. In [6], Residual CNN model [9, 10] with triplet loss function from face recognition community [11], was reported to perform much better than i-vector system training with 50k speakers. From the results above, it seems that end-to-end systems with speaker embedding outperforms i-vector systems on short duration conditions.

In this paper, we propose a novel end-to-end system for text-independent speaker identification on short utterances. Firstly, we employ a GRU network to extract temporal utterance-level feature, which is expected to retain one's speaking style. Then, a ResCNN model is trained to make robust speaker embedding. Finally, the whole system, including the GRU and ResCNN model, is jointly optimized by using the proposed loss function, called speaker identity subspace loss. The novel loss assumes that one's utterances in various conditions can be mapped to a single ideal object in the latent space. Specifically, as one regularization term, the local consistency constraint is incorporated into the proposed loss function. The motivation behind is that the distribution of the latent ideal objects should preserve the geometric structure of the obtained voice space. To handle the training difficulty, we utilize the recent advancements in deep learninig community such as batch normalizaition and network reduction [12]. We test our proposed system on speaker identification task with a Short Duration Corpus. Experiments show that GRU learned feature contains more speaker characteristic, and the speaker identity subspace loss significantly improves the discriminative ability of our system.

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents the end-to-end approach and speaker identity subspace loss. The performance of our proposed system is evaluated, and the results are discussed in Section 4. Section 5 gives a conclusion of this paper.

2. Related work

A great number of neural networks have been applied to learning feature. DNN manages to preserve the task-related information by the layer-by layer processing [13, 14]. CNN models spectral correlations in acoustic features through treating the feature as a 2-D image [15]. GRU captures the sequential nature of audio signals using purpose-built memory cells to find and exploit long range information [6]. Though these neural networks made great achievements, they still have some shortcomings. In general, DNN doesn't use any information about the phone content, while CNN thinks little of the temporal constraint, and the output of GRU is always averaged to get utter-level embeddings, ignoring one's speaking style. To cope with

The research work is supported by the National Key Research and Development Program of China under Grant No. 2016YFB1001404.

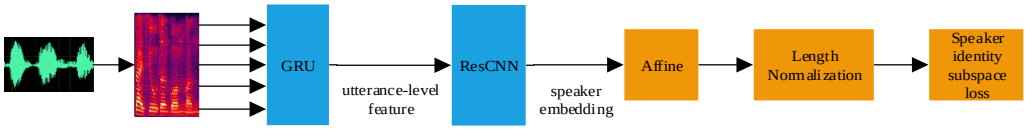

Figure 1: *Architecture of Our Proposed Speaker Identification System*

this problem, we only connect the last output of GRU to the affine layer to obtain a single utterance-level feature, and further employ ResCNN to learn discriminative speaker embeddings. In this way, not only the sequential nature is captured, but also the spectral correlations is well modeled in the embeddings.

Many distance-based loss, such as PLDA-based loss [3], contrast loss [16] and triplet loss [17], have been introduced to optimize models. Among them, PLDA-based loss compares pairs of embeddings with a multiclass cross entropy objective, while contrast loss focuses on difficult pairs of embedding vector and negative centroid, and triplet loss calculates the similarity of three samples that two from the same person and one from another. They have similar goal to maximize between-class distance while minimizing within-class distance. On the contrary, we focus on finding the unique ideal object in the latent subspace that one's utterances map to, and put forward a novel loss function, called speaker identity subspace loss. Its efficiency and discriminative ability are showed in the experiments.

3. Model and Approach

Our proposed architecture is presented in Fig.1. Utterances are first processed to frame-level features, and further transformed to utterance-level sequential feature by GRU. Then, ResCNN learns robust speaker embedding together with an affine layer and a length normalization layer, as detailed in Section 3.2. Finally, the speaker identity subspace loss layer estimates the identity of the speaker embeddings, as described in Section 3.3.

3.1. GRU Network Based Feature Learning

Recurrent networks have shown great success in speech recognition [18] in frame-level feature extraction. GRU [19, 20], an extension of LSTM, has simpler structure and smaller parameters. Although GRU is not more accurate than LSTM, it runs faster and is less likely to diverge. Therefore, we use GRU to research temporal feature.

The model hyperparameters of proposed GRU is showed in Table 1. For faster GRU layer computation, a 5×5 filter size, 2×2 stride convolution layer is applied, helping to reduce dimensionality in both time and frequency domains. Following the convolutional layer is a unidirectional GRU layer with 1024 units. Different from the traditional GRUs averaging multiple outputs [6], we only connect the last output to the affine layer to obtain a single utterance-level feature. Then, the affine layer is applied to adjust the feature dimension to the input size of ResCNN. Besides, to reduce internal covariate shift, we use sequence-wise batch normalization and clipped-ReLU activation [20] in the model.

3.2. Residual CNN Speaker Embedding Model

CNNs have also achieved great performance on speaker recognition [21]. However, as the network gets deeper, training grows rather difficult. ResNet [11], composed of a number of stacked residual blocks (ResBlocks), is reported to work well in easing training. By introducing the residual connections to the CNN network, our model achieves faster convergence without adding additional computation complexity.

Table 1: *Architecture of the GRU model*

layer name	struct	stride	dim	param
conv16-s	5×5,16	2×2	2048	1.2 K
GRU	1024 cells	1	1024	7.8 M
affine	1024×3072	-	3072	3.2 M
ln	-	-	512	-
total	-	-	-	11.1 M

Table 2 describes the structure of ResCNN. The input utterance-level feature is first reshaped to $32 \times 32 \times 3$, treated as a 3-d image to the ResCNN network. After going through a 5×5 filter size, 2×2 stride convolution layer, the reshaped feature is further fed into three stacked ResBlocks. Each of the ResBlocks contains two convolutional layers with 3×3 filters and 1×1 strides, which helps to dig more useful information while saving resource. Then, a convolution layer and three stacked ResBlocks are employed. Moreover, we also adopt sequence-wise batch normalization (BN) between the convolution layer and nonlinear layer.

Table 2: *Architecture of the ResCNN model*

layer name	struct	stride	dim	param
reshape	$3072 \rightarrow 32 \times 32 \times 3$	-	3072	-
conv16-s	5×5,16	2×2	3072	1.2K
res16	$[(3 \times 3, 16) \times 2] \times 3$	1×1	3072	$2.4K \times 6$
conv32-s	5×5,32	2×2	3072	13K
res32	$[(3 \times 3, 32) \times 2] \times 3$	1×1	3072	$9.4K \times 6$
affine	3072×512	-	512	1.6M
ln	-	-	512	-
novel loss	-	-	512	-
total	-	-	-	10.1 M

3.3. Speaker Identity Subspace Loss

3.3.1. Basic Loss Formulation

In real scenarios, one's utterances may be captured in various conditions, like various emotional states and diverse channels. Sometimes, it turns rather hard for the identification task as the condition changes. Similar problem is encountered in the face recognition field, that the change of pose makes the identification of the face image difficult. Many cross pose recognition models have been put forward, such as Multi-view Discriminative Analysis (MvDA) [22], Tied Factor Analysis (TFA) [23], and achieved great improvements. The common goal of these existing approaches is to build a bridge between the observed image space and the pose free representation space. Inspired by this, we assume that one's utterances captured on different conditions, can be viewed as transformations of a single ideal object. Then, the transformation is described as:

$$h_i = \Gamma(\theta, x_{ij}) + \varepsilon_{ij} \quad (1)$$

where h_i is the ideal identity vector of speaker i, and x_{ij} is the jth utterance of speaker i; Γ is the transfer function, and θ is the parameter. In our work, Γ denote the GRU and ResCNN networks, and θ represent the weights of the networks. ε_{ij} is the noise term, which represents the background environment, etc.

As the existence of the noise term ε_{ij}, we can just obtain the estimated identity vector $\widehat{h_e}$:

$$\widehat{h_e} = \Gamma(\theta, x_{ej}) \tag{2}$$

In our work, $\widehat{h_e}$ is the output of the ResCNN network, and h_i is initialized orthogonally and updates along with the network. For the training set, our goal is to seek for the parameters θ and h_i that make one's estimated identity vector as close as possible to his ideal identity vector. The motivation can be formulated in the cost function as:

$$\begin{aligned} \min_{h_i{}_{i=1}^{M},\theta} \quad & L_{basic} = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\|h_i - \Gamma(\theta, x_{ij})\|_2^2 \\ \text{subject to} \quad & \|h_i\|_2^2 = 1,\ i = 1, \ldots, M. \end{aligned} \tag{3}$$

where N is the number of utterances, M is the number of speakers, and n_i is the utterances number of speaker i.

However, the above approach is likely to go into overfitting and thus has poor generalization ability. To deal with this problem, the local consistency constraint is introduced to the objective function as one regularization term.

3.3.2. Local Consistency Constraint

Locality information is an essential clue in manifold learning [24]. In most mainfold learning methods, researchers manage to find a subspace that optimally preserves the local structure of the observed data. In our work, the idealized identity vector can be regarded as the low-dimensional embedding of his utterances, so the distance relation of the identity vectors should be kept in accordance with the local relation in the observed utterance space. Therefore, we achieve this goal by minimizing the following energy function:

$$C_{accordance} = \sum_{p=1}^{M}\sum_{q=1}^{M}\|h_p - h_q\|^2 R_{pq} \tag{4}$$

where h_p, h_q are the identity vectors of speaker p and q. R_{pq} is the voice distance relation of speaker p and q. We select one utterance x captured on the same condition from each speaker [25],and define R_{pq} as:

$$R_{pq} = \begin{cases} e^{\frac{-\|x_p - x_q\|_2^2}{d_p d_q}} & \text{if } x_p \in N_s(x_q) \text{ or } x_q \in N_s(x_p) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where d_p denotes the distance from x_p to its rth nearest neighbour, and $N_s(x_p)$ is the s nearest-neighbors of x_p. r and s are the local scale factors, which determine the relative contribution of local structure preserved in the observed space. Thus, the decreasing speed of R_{pq} is well controlled by the distance between x_p and x_q.

In summary, the speaker identity subspace loss is defined as follows:

$$\begin{aligned} & \theta, h_i{}_{i=1}^{M} = \arg\min L \\ & \text{subject to} \quad \|h_i\|_2^2 = 1,\ i = 1, \ldots, M. \end{aligned} \tag{6}$$

where L is the weighted sum of L_{basic} and $C_{accordance}$, and formulated as:

$$L = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\|h_i - \Gamma(\theta, x_{ij})\|_2^2 + \lambda\sum_{p=1}^{M}\sum_{q=1}^{M}\|h_p - h_q\|^2 R_{pq} \tag{7}$$

4. Experiments

In this section, we first present the database used in the experiments, and then report the results. All the experiments are conducted with the Keras toolkit [26].

4.1. Data Sets

The corpus used for the network training and evaluation, is mainly collected from three different channels, i.e., Interview, Android mobilephone, and Apple mobilephone. It comprises 968 speakers, 35,983 utterances, about 27 hours utterances in Mandarin. Each person has about 37 utterances, the durations of which are mostly around 2 to 5 seconds. The corpus is split into training and evaluation by randomly selecting 100 speakers to be the evaluating set. The details of the dataset are shown in Table 3.

Table 3: *Corpus statistics*

	#spk	#utt	#utt/spk	dur/utt
Training	868	32,322	37.2	2.74s
Eva	100	3,661	36.61	2.72s
Total	968	35,983	37.17	2.74s

4.2. Basic Setup

To observe the performance of GRU Network and speaker identity subspace loss, we build an averaged ResCNN model as the baseline. The raw audio is converted to 64-dimensional log mel-filter bank (Fbank) coefficients [2] with a frame-length of 25ms. This raw feature is augmented by its first and second order derivatives, and further splices the neighboring frames with a symmetric 8-frame window. Then, a frame-level energy-based Voice Activity Detector (VAD) selection and a utterance-level averaging operation are made to the feature. After that, the averaged feature is directly input to the ResCNN model described in Section 3.2.

For our proposed end-to-end model, raw audio is converted to 64-dimensional Fbank coefficients, further augmented by its first and second order derivatives and frame-level energy-based VAD selection. Then, the feature is input to the GRU model described in Section 3.1, and the output is further input to the ResCNN model described in Section 3.2.

It is found that pre-training the model using a softmax layer and cross entropy loss over a fixed list of speakers achieves great improvements. Therefore, the models are trained in two stages: softmax pre-training and speaker identity subspace loss fine-tuning. In both stages, we use ADAM [27], with a linear decreasing learning rate from 0.05 to 0.001. In the fine-tuning stage, the weight ratio between softmax loss and speaker identity subspace loss ranges from 10 to 0.1.

4.3. Averaged Feature vs. Time Sequential Feature

Averaged feature is the feature learned by the baseline, while time sequential feature is the GRU learned feature. To investigate the performace of GRU Network, we compare the performance of the two features under the same conditions. Partial

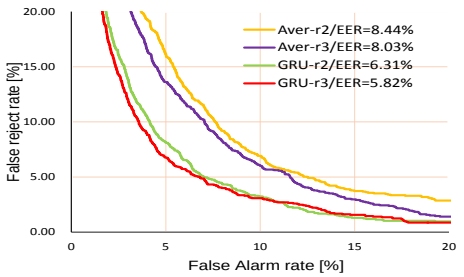

Figure 2: *DET plots of the averaged and GRU learned feature. r2 denotes that the learning rate is* 10^{-2}*, Aver means that the model is fed with the averaged feature.*

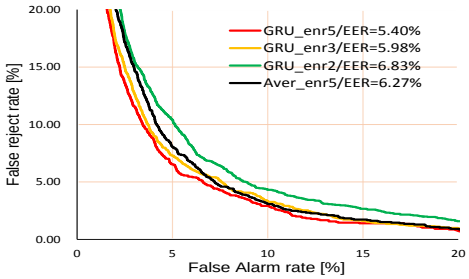

Figure 3: *DET plots on different enrolled utterances number. Models are trained with the speaker identity subspace loss in the learning rate of* 10^{-3}*.* $enr*$ *denotes that the model is trained with* $*$ *utterances.*

results are showed in Fig.2. It is gratifying to see that GRU learned feature significantly promotes the performance of network against averaged feature. In the term of EER, the amount of relative reduction are 25.23% and 27.52% respectively, in the learning rate of 10^{-2} and 10^{-3} correspondingly. This may be largely due to the powerful ability of GRU to capture the sequential nature, and the convenience brought by the operation selecting the last output of GRU in obtaining utterance-level feature. Also, it verifies the assumption that speaking style is an vital clue in distinguishing speakers, which can be well retained by GRU. In addition, the learning rate also makes slight effect on the performance, that the smaller one obtains better result. In the following experiments, we set the learning rate to 10^{-3}.

4.4. Speaker Identity Subspace Loss Fine-tuning

Speaker identity subspace loss pays much attention on minimizing the within-class distance, but is inferior to softmax loss in maximizing the between-class distance. Therefore, we adopt the joint supervision of softmax loss and speaker identity subspace loss in the fun-tuning stage, whose results are shown in Table 4. As we can see, the weight ratio between softmax loss and speaker identity subspace loss makes difference on the system performance. When it turns to 1:2, our proposed system and baseline system achieve their best performance, with 7.22% and 21.92% relative improvement obtained respectively. It indicates that our proposed loss has more advantages in digging discriminative information in the feature, and the weight ratio between the two loss functions should be balanced well.

Table 4: *Speaker identity subspace loss against softmax loss.* $GRU/(1:5)$ *means the weight ratio between softmax loss and speaker identity subspace loss is 1:5 in our end-to-end system.*

system	Aver	GRU/(1:10)	GRU/(1:5)	GRU/(1:2)	GRU/(1:1)	GRU/(1:0.1)
sof	0.0803	0.0582	0.0582	0.0582	0.0582	0.0582
sof+sis	0.0627	0.0587	0.0571	**0.0540**	0.0568	0.0612

4.5. Performance by Number of Enrolled Utterances

In the speaker identification, there is always a vital step that builds the speaker model with enrolled utterances. In our work, we average embeddings across one's enrollment utterances to make his final representation. Take into account the effect of enrolled utterances number, we made experiments and the results are present in Fig.3. A DET curve for the averaged feature learned baseline is also illustrated for system comparison. In Fig.3, a performance degradation is observed when we reduce the enrolled number from 5 to 2. In terms of EER, the amounts of relative degradation are 10.74% and 14.21% when the number of enrolled utterances decreases from 5 to 3, and 3 to 2. The observation reminds us that the number of enrolled utterances is an essential item in designing speaker verification systems, and 3 to 5 utterances may be appropriate.

4.6. Performance against Different Duration

The system is further applied to different duration conditions, and the performances are indicated in Table 5. More specifically, 2s, 3s, 5s, and 8s conditions are tested with our end-to-end framework. Sharp increasements are witnessed when the duration grows from 3s to 5s, and 5s to 8s, whose relative improvements are 32.57% and 22.32%. Mainly because that longer utterances contain more speaking style information and spectral correlations, which benefits the discriminative speaker embeddings learning. The observation shows guidance to us that 5s is enough in training short-duration speaker identification system. Plus, the longer duration may bring further improvement as well as redundancy. It is also found that, network training with speaker identity subspace loss works better than softmax loss in the three duration periods, further confirming the experment in Section 4.4.

Table 5: *System performance on different duration conditions*

system	Aver/2s	GRU/2s	GRU/3s	GRU/5s	GRU/8s
sof	0.0803	0.0582	0.0560	0.0406	0.0352
sof+sis	0.0627	0.0540	0.0525	0.0354	**0.0275**

5. Conclusions

In this study, we present a novel end-to-end text-independent speaker identification system. GRU allows us to learning one's speaking style, and ResCNN is successfully applied to modeling speaker discriminative embedding. Speaker identity subspace loss based on the Euclidean distance makes it possible to optimize the entire system jointly. Experiments show that our end-to-end system achieves consistently better performance than the previous methods, and the speaker identity subspace loss makes the embeddings more discriminative.

It is believed that our system can be extended to related areas, such as speaker verification, diarization and clustering. Of course, there are still some improvements to make through different loss functions, front feature processing, neural networks, etc. Additionally, the growing size of model along with the increased data amount, is also a problem we have got to consider. All in all, this system shows both decent advancements and a direction where our further research goes forward.

6. References

[1] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[2] N. Dehak, P. J. 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.

[3] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech 2017*, pp. 999–1003, 2017.

[4] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 165–170.

[5] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," *CoRR*, vol. abs/1509.08062, 2015. [Online]. Available: http://arxiv.org/abs/1509.08062

[6] 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.

[7] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using cnns." in *INTERSPEECH*, 2016, pp. 3434–3438.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[10] 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.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[13] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[14] T. Yao, C. Meng, H. Liang, and L. Jia, "Speaker recognition system based on deep neural networks and bottleneck features," *Journal of Tsinghua University (Science and Technology)*, vol. 56, no. 11, pp. 1143–1148, 2016.

[15] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *arXiv preprint arXiv:1701.02720*, 2017.

[16] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *arXiv preprint arXiv:1710.10467*, 2017.

[17] C. Zhang and K. Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," in *Proc. of Interspeech*, 2017.

[18] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[20] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[21] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 171–178.

[22] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[23] S. J. Prince, J. H. Elder, J. Warrell, and F. M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 6, pp. 970–984, 2008.

[24] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[25] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in neural information processing systems*, 2005, pp. 1601–1608.

[26] F. Chollet *et al.*, "Keras," 2015.

[27] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.