MULTI-TASK KNOWLEDGE DISTILLATION WITH RHYTHM FEATURES FOR SPEAKER VERIFICATION

Ruyun Li[1], Peng Ouyang[2], Dandan Song[2] and Shaojun Wei[1]

[1]Department of Microelectronics and Nanoelectronics, Tsinghua University, Beijing, China
[2]TsingMicro Co. Ltd., Beijing, China

ABSTRACT

Recently, speaker embedding extracted by deep neural networks (DNN) has performed well in speaker verification (SV). However, it is sensitive to different scenarios, and it is too computationally intensive to be deployed on portable devices. In this paper, we first combine rhythm and MFCC features to improve the robustness of speaker verification. The rhythm feature can reflect the distribution of phonemes and help reduce the average error rate (EER) in speaker verification, especially in intra-speaker verification. In addition, we propose a multi-task knowledge distillation architecture that transfers the embedding-level and label-level knowledge of a well-trained large teacher to a highly compact student network. The results show that rhythm features and multi-task knowledge distillation significantly improve the performance of the student network. In the ultra-short duration scenario, using only 14.9% of the parameters in the teacher network, the student network can even achieve a relative EER reduction of 32%.

KEYWORDS

Multi-task learning, Knowledge distillation, Rhythm variation, Angular softmax, Speaker verification

1. INTRODUCTION

Using Deep Neural Network (DNN) to extract speaker embeddings has shown impressive performance in speaker verification (SV). Speaker embeddings denote fixed-dimensional vector-based representations for modeling the characteristics of speakers.

Gaussian Mixture Network-Universal Background Network (GMM-UBM) system dominated the SV field for one decade since proposed in [1]. Inspired by Joint Factor Analysis in [2], i-vector [3] was proposed and represented the state-of-the-art speaker networking framework. Recently, speaker embeddings [4, 5, 6, 7] learning with DNN has become mainstream for speaker networking in SV. By averaging the frame-level extracted deep features, the segment-level representation of a recording is obtained, which is called d-vector [8]. Some researchers follow and extend this work by replacing the simple neural network with complicated architectures such as Convolutional Neural Network (CNN) and Time-Delay Neural Network (TDNN) or redesign the optimization metric and propose new embeddings such as j-vector [9]. Instead of training the

David C. Wyld et al. (Eds): ITCSE, NLCA, ICAIT, CAIML, ICDIPV, CRYPIS, WiMo - 2020
pp. 249-262, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100523

DNN on the frame level, researchers in [10] add a temporal pooling layer and train the network on the segment level, which is called x-vector. It is proven to achieve excellent performance.

Advanced loss functions also benefit to build a more powerful deep architecture, such as triplet loss [6], the generalized end-to-end loss [11], and the angular softmax [5]. The angular softmax (A-softmax) modifies the softmax loss function to learn angularly discriminative embeddings and adds a controllable parameter to pose constraints on the intra-speaker variation of the learned embedding.

Even if the methodology above reported impressive low error rates (≈1\% [3]), SV is still challenging in different trial conditions and linguistic environments like diverse phonological content. Besides, x-vector is too computationally intensive to be deployed on portable devices.

Among the efforts to compress these networks, knowledge distillation is a natural method, where a large network (teacher) provides weighted targets to guide the training of a small network (student). However, previous studies only explored the effect of single-level knowledge distillation on speaker embedding performance, and single-level knowledge distillation was not effective enough to obtain highly compact networks with better performance than large networks. In this paper, phonological content is considered in extracting speaker acoustic features to improve the performance of intra-speaker verification. We calculate seven rhythmic parameters, which is based on temporal characteristics of speech intervals. Then we concatenate these rhythm features with MFCC features. Besides, we aim to build small networks that need much fewer resources and are more suitable for deployment without performance degradation. Multi-task knowledge distillation utilizes the embedding-level and label-level output of teacher networks [12] to guide the training of student networks, to reduce the performance gap between student networks and teacher networks. Sometimes, student networks even outperform teacher networks, because of the dark knowledge [13] transferred in distillation.

The main contributions of this article are as follows:

1. The fusion of rhythm feature and MFCC feature: rhythm parameters are multiplied by a weight factor, and then concatenated with the MFCC feature.

2. Multi-task knowledge distillation: The main task is to force the student network to emit posterior probabilities similar to the hard speaker labels. Besides, we utilize label-level and embedded-level knowledge distillation to guide the training of the highly compact student network as two auxiliary tasks. The total loss comes from these three tasks.

3. A highly compact student network has competitive performance with the teacher network: In the ultra-short duration scenario, a student network can even achieve a 7.02% relative EER reduction, using only 13% of the parameters in the teacher network. We studied the effects of deep speaker embedding, A-softmax, rhythm features, and multi-task knowledge distillation on intra- and inter-speaker verification (false miss and false alarm), which shed light on the success of our methods.

The rest of the article is organized as follows. Section 2 briefly introduces deep speaker embedding learning. Section 3 introduces the multi-task knowledge extraction with rhythm function. Sections 4 and 5 show the experimental setup and results, respectively. Section 6 summarizes the paper.

2. RELATED WORK

2.1. Deep Speaker Embedding Learning

Deep speaker embedding learning has been dominating the field of text-independent speaker verification. Powered by advanced computational resources, and large-scale speech datasets, e.g., VoxCeleb and speaker verification corpora packaged by the National Institute of Standards and Technology (NIST), it is possible to train very deep networks to extract speaker embeddings (segment-level representations).

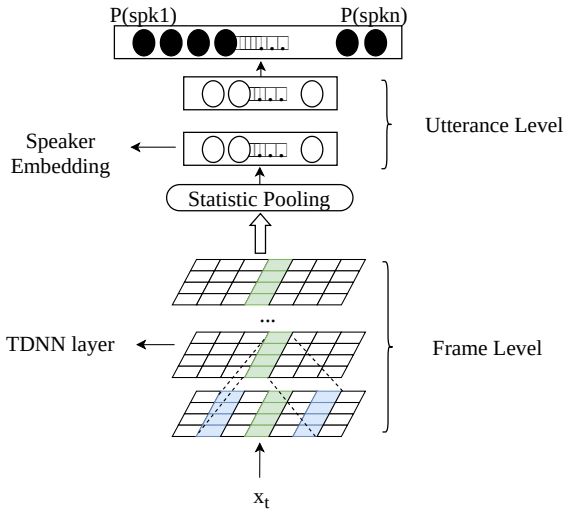


Figure 1 Network architecture of x-vector

In this paper, we adopt the normal x-vector architecture. The DNN used in the x-vector is depicted in Figure 1, and the detailed network configuration is described in Section 4.1.1. In our work, the pooling mechanism calculates the mean and standard deviation of the frame-level representations, but several studies have extended it to multi-head attention layers [20, 21] and learnable dictionary layers. We use angular softmax loss (A-softmax) as the training criterion, which was proposed for face recognition and introduced to speaker verification in [22, 23]. The back-end technology is cosine distance scoring and probabilistic linear discriminant analysis (PLDA) [24, 25].

2.2. Knowledge Distillation

There have been efforts to compress these networks, e.g., parameter pruning and sharing [26], low-rank factorization [27] and knowledge distillation [29, 30]. Knowledge distillation has proven a promising way to narrow down the performance gap between a small network (student) and a large network (teacher). It works by adding a term to the usual classification loss, which encourages students to imitate the behavior of teachers. However, knowledge distillation for deep speaker embedding has not been investigated thoroughly in the literature. [29] built a distillation framework to learn the distribution of embedding vectors directly from the teacher. [30] further recommends using the teacher's speaker posterior probability as reference labels for the student. These previous studies only explored the effect of single-layer knowledge distillation on speaker embedding performance, and single-layer knowledge distillation was not effective enough to obtain highly compact networks with better performance than large networks.

3. MULTI-TASK KNOWLEDGE DISTILLATION FOR DEEP SPEAKER EMBEDDING

This section describes the speaker verification systems developed for this study, which consist of fusion features, and multi-task knowledge distillation. Our experiments are based on the Kaldi speech verification toolkit [14].

3.1. Fusion of Rhythm Features and MFCC Feature

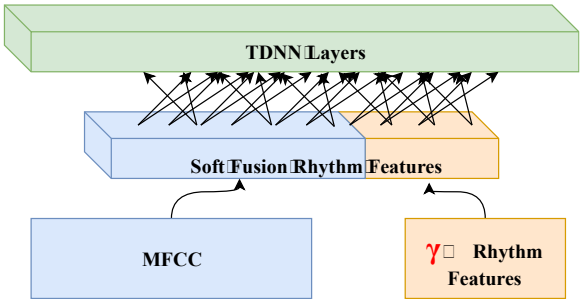


Figure 2 Fusion of rhythm features and MFCC feature

Rhythm variations are proven to have virtually a significant impact on intra-speaker verification, which are commonly used in the field of speech rhythm research [15, 16, 17, 18]. In this paper, we introduce seven rhythm variation measurements to improve speaker verification performance: $\%VO$, $\overline{VO}$, $VarcoUV$, $VarcoVO$, $\%(UV_{i+1} > VO_i)$, $Average(pair)$ and $VarcoPair$, which are formulated in [19].

In the x-vector framework, silenced frames are filtered out by voice activity detection (VAD). As shown in Figure 2, we multiply rhythm features with a weight factor g, then combine them with MFCC feature. Our rhythm variation measurements are based on voiced and unvoiced durations (including pauses), which is detected via python interface of the WebRTC VAD. Our experiment in Section 5.1 will investigate the best value of g for fusion.

3.2. Multi-task Knowledge Distillation

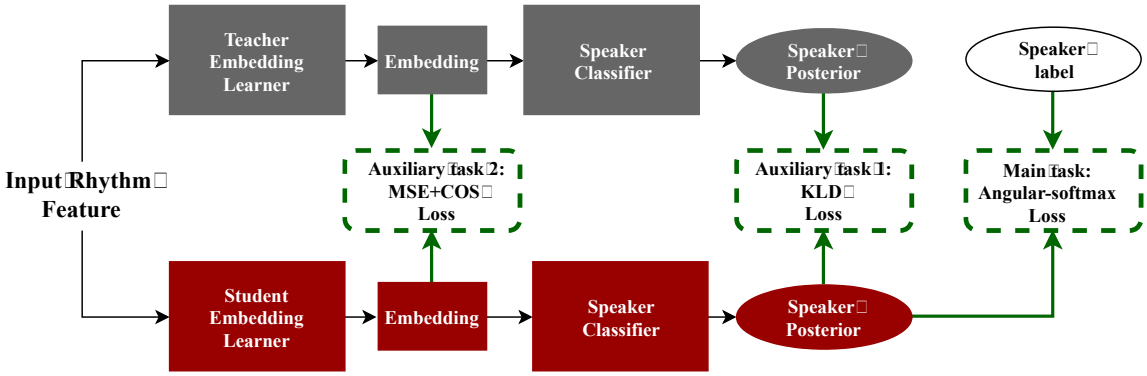


Figure 3 Multi-task knowledge distillation architecture: The system consists of three parts, the teacher network (in Grey), the student network (in Red), and three training tasks.

Multi-task knowledge distillation forces the student network to train on multiple different, but related knowledge distillation tasks, which can make better use of the teacher network. As shown in Figure 3, the multi-task knowledge distillation for deep speaker embedding network includes three tasks:

1. The main task (see Equation 1) is to train the student speaker embedding network over the same set of speakers directly with hard speaker labels, just like what we did for a teacher network.

2. Label-level knowledge distillation (see Equation 3), where the optimization of the student network is guided by the posteriors predicted by a well pre-trained teacher network.

3. Embedding-level knowledge distillation (see Equation 4), which directly constrains the similarity of speaker embeddings learned from the teacher and student network.

For the main task, instead of using the categorical cross-entropy for training, we use the A-softmax loss for classification. A-softmax has more stringent requirements for correct classification when $m \geq 2$ (an integer that controls the angular margin), which generates an angular classification margin between embeddings of different classes. The A-softmax loss is formulated in Equation 1:

$$L_{A\text{-}softmax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{\|x_i\|\psi(\theta_{y_i,i})}}{Z} \quad (1)$$

$$Z = e^{\|x_i\|\psi\theta_{j,i}} + \sum_{j=1, j\neq y_i}^{N} e^{\|x_i\|\cos(\theta_{j,i})}, \psi(\theta_{y_i,i}) = (-1)^k cos(m\theta_{y_i,i}) - 2k, \quad (2)$$

where N is the number of training samples; x_i is the input of the last (i.e. output) layer; y_i is the ground truth label for the i^{th} sample; w_j is the j^{th} column of the weights in the output layer; $\theta_{j,i}$ is the angle between w_j and x_i; $\theta_{y_i,i} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$ and $k \in [0, m-1], m \geq 0$.

We distill the knowledge from the output of label-layer. Label-level knowledge distillation means the optimization of student network is guided by the posteriors predicted by a pre- trained teacher network. The objective is defined as:

$$L_{KLD} = -\sum_{i=1}^{N}\sum_{j=1}^{C} \tilde{y}_j^i \log y_j^i, \quad (3)$$

where C is the number of speakers in the training set; $\tilde{y}^i$ is the posteriors of the i-th sample predicted by the teacher network. Other definition of symbols is the same as Equation 1.

In addition to the label-level knowledge distillation, Assuming the student and teacher produce the same dimension of speaker embeddings, embedding-level knowledge distillation directly constrains the similarity of speaker embeddings learned from the teacher and student network, which is formulated as:

$$L_{COS} = -\sum_{i=1}^{N} \frac{v_i^t v_i^s}{\|v_i^t\| \|v_i^s\|}, \quad (4)$$

where v_i^t represents the embedding computed by the teacher network for the i^{th} sample; v_i^s denotes the embedding computed by the student network.

In the optimization, losses of these three tasks are combined to train the student network as:

$$L_{total} = L_{A\text{-}softmax} + aL_{KLD} + bL_{COS}, \qquad (5)$$

where a and b are hyper-parameters to balance three losses.

4. EXPERIMENTAL SETUP

4.1. Dataset

We evaluate the performance of our method on a short duration text-independent dataset called XiaoAi-Speech, which consists of 230288 clean utterances from 448 male individuals. Each utterance varies between 1 and 8 seconds (before removing the silenced frames). The database contains almost 320h recordings. It is mainly used for short-duration speech processing as it contains relatively short-duration phrases. Besides, it allows studies on intra-speaker and inter-speaker comparisons, because each speaker provides nearly 500 utterances of different content. We report the speaker verification results on this dataset in ultra-short-duration, short-duration, and normal-duration scenarios, respectively.

4.1.1. Training Data

The training data contains 248 speakers, and each speaker has almost 500 utterances.

The i-vector extractor is a 2048 component GMM-UBM, which is trained on full-length recordings using 23-dimensional MFCC speech features. Short duration i-vectors and ultra-short duration i-vectors are extracted from the first 10n frames of the test data, where n is the duration of recordings being considered (in ms). We only choose speakers that have more than 8 recordings (with 3~5-second durations).

4.1.2. Evaluation Data

We focus on the case where both the enrollment and test recordings of a verification trial are in the same duration scenario. In the normal-duration scenario, 9 enrollment and 2 test utterances are prepared for each speaker. In the short-duration scenario, 7 enrollment and 2 test utterances are prepared. In the ultra-short-duration scenario, 2 enrollment and 2 test utterances are prepared. The evaluation set is selected from our database, and there is no speaker overlap with the training set. The enrollment part contains 200 speaker models, and the test part contains 800 utterances from the 200 models in the enrollment set, of which 400 are for ultra-short-duration trials, and the rest 400 are for short-duration trials. There are 20K trials in the entire trial list, including 50% intra-speaker (target) trials and 50% inter-speaker (non-target) trials.

4.2. Speaker Verification

In this section, we present our experimental setup, as well as details related to input features, neural network training, and classifiers. All systems compared in this paper are presented in Table 1. All deep speaker embeddings systems in this paper are trained on 23-dimensional MFCC

features (sometimes combined with 7-dim rhythm variation features) with a frame-length of 25ms that are mean-normalized over a sliding window of up to 3 seconds of short-duration snippets speech. Our experiments are conducted in three evaluation scenarios: Short-duration (3~5-second), Ultra-short-duration (1~3-second) and Normal-duration (1~5-second) evaluation recordings.

4.2.1. Large-scale System

The I-vector system relies on a universal background network and a total variability matrix, which is called *i-vec*. Input features are 23-dimensional MFCC with first and second-order time derivatives. The number of Gaussian components is set to 2048, while the dimension of the i-vector is 600.

Table 1 The configuration of our systems

Network	I-vector dim	X-vector			#Model size
		#Input	#TDNN layers	#Neurons	
i-vec	600	23	n/a	n/a	20.38M
x-vec	n/a	23	3	512	18.3M
x-vec+asoftmax	n/a	23	3	512	18.3M
x-vec+asoftmax+rhythm	n/a	30	3	512	23.15M
student-64	n/a	30	3	64	3.45M
student-32	n/a	30	3	32	763K

The x-vec systems, x-vector+asoftmax system, and x-vec+asoftmax+rhythm system are described in Section 2 and Section 3.1, respectively. We adopt x-vec+asoftmax+rhythm, an x-vector based on A-softmax and rhythm features, as the teacher network in the following multi-task knowledge distillation experiments, since an excellent performance was reported using this architecture on XiaoAi-Speech. The detailed network configuration of the teacher network is shown in Table 1. The input acoustic features are fed into an eight-layer DNN. The first five layers Frame 1 to Frame 5 are constructed with a frame-level time-delay architecture. The statistics pooling layer aggregates over frame-level output vectors of the DNN and computes their mean and standard deviation. This pooling mechanism enables the DNN to produce fixed-length representation from variable-length speech segments. Then their mean and standard deviation are concatenated together and forwarded to two additional hidden layers segment 6 and segment 7. Finally, the system is optimized using stochastic gradient descent (SGD) using A-softmax. The N on the A-softmax layer corresponds to the number of training speakers. We also decayed the learning rate every 4 epochs. During the inference phase, speaker embeddings are extracted from the affine component of layer segment 6 before the nonlinearity. Then a PLDA backend is used to compare pairs speaker embeddings.

Table 2 The architecture of *x-vec+asoftmax+rhythm*.

Layer	Layer context	Total context	Input x output	#Parameter
frame1	[t-2, t+2]	5	30x512	30x5x512
frame2	[t-2, t+2]	9	512x512	512x5x512
frame3	[t-3, t+3]	15	512x512	512x7x512
frame4	{t}	15	512x512	512x512
frame5	{t}	15	512x1500	512x1500

stats pooling	[0, T)	T	1500xTx3000	0
segment6	{0}	T	3000x512	3000x512
segment7	{0}	T	512x512	512x512
A-softmax	{0}	T	512xN	512xN

4.2.2. Small-scale System

Several different setups for highly compact student networks are investigated in our experiments. The most natural choice is to use a shallower x-vector. Two setups are adopted, namely *student-64* and *student-32*, with the number of hidden units for TDNN layers set as 64 and 32, respectively. Both teacher and student networks are of the same 512 speaker embedding dimension. During the inference phase, the student network was used to predict speaker embedding vectors for enrolment and test data, which was then followed by PLDA scoring, which is the same as the teacher network.

4.3. Evaluation metric

To further investigate the impact of out methods on intra- and inter-speaker verification separately, we use C_{llr} instead of hard decision like equal error rate (EER) to evaluate the log-likelihood-ratio (LR) of speaker pairs. C_{llr} can evaluate the discriminant ability of the log-likelihood ratio (LR) of the speaker pair, while EER is valid for overall correct-classification rate. C_{llr} is calculated as followed:

$$C_{llr} = \frac{1}{2N_{tar}} \sum_{LR \in \mathcal{X}_{tar}} log_2(1+\frac{1}{LR}) + \frac{1}{2N_{non}} \sum_{LR \in \mathcal{X}_{non}} log_2(1+LR) \quad (6)$$

As shown in Equation 6, C_{llr}^{TAR} is the average information loss corresponding to target trials, while C_{llr}^{NON} is the average information loss corresponding to non-target trials. C_{llr} is the sum of the two parts. The lower the C_{llr}, the better the performance is.

5. RESULT AND ANALYSIS

5.1. Fusion of Rhythm Features and MFCC Feature

Based on the *x-vec+asoftmax+rhythm* system, we optimize the weight parameter g for feature fusion to minimize EER. Figure 4 shows the EER on the corresponding evaluation set under ultra-short duration, short duration, and normal duration scenarios, respectively. The results motivated us to choose $g = 0.01$ for feature fusion, which produced the lowest EER in all scenarios.

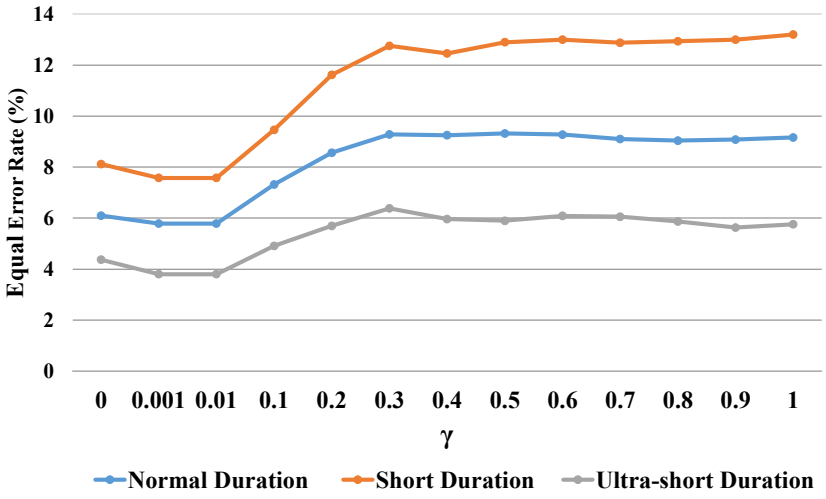


Figure 4 Optimization of the weight factor g in feature fusion

5.2. Effect on Intra- and Inter-speaker Verification

As shown in Table 2, rhythm features and multi-task knowledge distillation can both improve intra-speaker verification (target comparisons) and inter-speaker verification (non-target comparisons). Although *i-vec*, *x-vec*, *x-vec+asoftmax*, *x-vec+asoftmax-rhythm*, and *student-64-TS* achieve lower and lower overall error rates (EER), their effects on the target and non-target comparisons are different from each other. Compared with *i-vec*, *x-vec* reduces EER by 7.5% at the cost of non-target comparison accuracy. Compared with *x-vec*, *x-vec+asoftmax* achieves a 4.6% EER reduction at the cost of target comparison accuracy. At the meantime, it is worth noting that *x-vec+asoftmax+rhythm* have significantly improved the target comparison, and it does not affect the accuracy of the non-target comparison, which is consistent with the conclusion in [19]. Besides, multi-task knowledge distillation makes the *student-64-TS* network have better performance in both target and non-target comparisons, with a 7.1% EER reduction. The results show that DNN is powerful in modeling high-dimensional speaker embedding, but under non-discriminatory training conditions, the performance of target comparison is worse than i-vector. A-softmax is more strict than conventional softmax, it imposes a larger angle margin between the speakers, and classifies the samples into the corresponding categories. Therefore, *x-vec+asoftmax* is reasonable to harm the accuracy of the target comparison.

Table 2 C_{llr}^{TAR} and C_{llr}^{NON} for different speaker verification networks. PLDA is the scoring back-end for EER.

Network	C_{llr}^{TAR}	C_{llr}^{NON}	EER (%)
i-vec	9.49	0.01	18.2
x-vec	1.64	0.16	10.7
x-vec+asoftmax	5.14	0.02	6.1
x-vec+asoftmax+rhythm	2.89	0.03	5.7
student-64-TS	1.41	0.03	3.6

5.3. Multi-task Knowledge Distillation

Table 2 EER and parameter comparison of different speaker embedding architectures. PLDA as the scoring back-end. TS denotes multi-task teacher-student learning. Compression ratio (CR) is the relative reduction rate of model size.

Network	TS	EER (%)			Model size	Compression Ratio
		Ultra-short	Short	Normal		
i-vec	No	19.15	13.19	15.8	20.38M	-
x-vec	No	13.04	8.86	10.67	18.2M	-
x-vec+asoftmax	No	8.12	4.37	6.1	18.3M	-
x-vec+asoftmax +rhythm	No	**7.38**	**3.78**	**5.73**	**23.15M**	-
student-64	No	11.62	5.69	8.57	3.45M	85.1
student-64-TS	Yes	**5.02**	**1.75**	**3.64**	**3.45M**	**85.1**
student-32	No	19.15	13.19	15.8	673K	97.1
student-32-TS	Yes	**7.195**	**3.505**	**5.74**	**673K**	**97.1**

As shown in Table 3, compared with student baselines, multi-task knowledge distillation significantly boosts the performance of student networks, and it could obtain highly compact networks with better performance than large networks. *X-vec+asoftmax+rhythm* is the teacher network, while *student-32* and *student-64* with no knowledge distillation are two student network baselines. *Student-32-TS* can achieve competitive performance with the teacher network by using only 2.9% of parameters used in the teacher. *Student-64-TS* can achieve a 32% relative EER reduction by using only 14.9% of parameters used in the teacher. The more compact the student network, the more significant the effect of multi-task knowledge distillation.

Figure 5 shows the DET test curves of different systems under normal duration conditions. *Student-32-TS* is competitive with *x-vec+asoftmax+rhythm* in all operating areas.

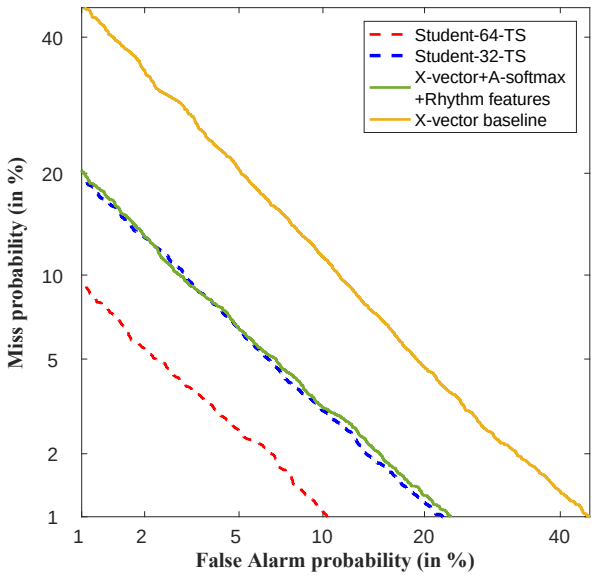
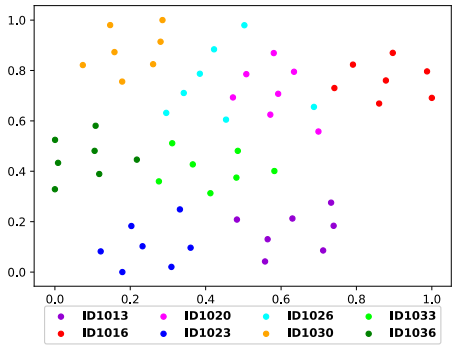


Figure 5 DET curve for baseline, teacher and two high compact student networks.

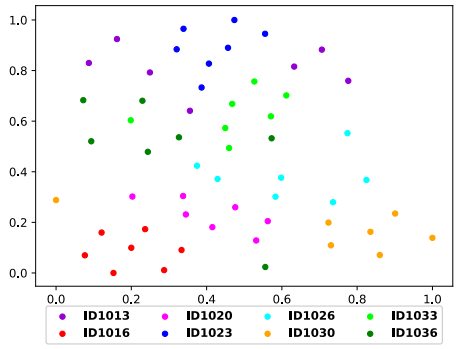
Figure 6 shows the t-SNE [30] plots corresponding to *x-vec+asoftmax+rhythm*, *student32*, and *student-32-TS* networks. The distribution of speaker embeddings in the 2D projected t-SNE space generally revealed speaker clusters. Eight speakers from the evaluation dataset were randomly selected, and for each, we use seven recordings spoken by the speaker (56 in total). The selected 56 samples were plotted in the 2D projected t-SNE space, with colors denoting different speakers.

Multi-task knowledge distillation can effectively pull intra-speaker samples closer and push inter-speaker samples further. On the one hand, as shown in Figure 6(b) and (c), compared with the *student-32* baseline, data points from the same speaker tend to be closer, while those from different speakers become more distinct. For instance, the distribution of x-vectors from speaker ID1013 (violet), ID1023 (blue) become denser. On the other hand, as shown in Figure 6(a) and (c), compared with the *x-vec+asoftmax+rhythm*, speaker subset clusters emerge in the student. Samples from ID1013 (violet), ID1023 (blue), ID1033 (lime) and ID1036 (green) formed a cluster while the rest formed another. Within each subset cluster, the student and teacher have a similar relative position of speaker clusters in the embedding space. X-vectors from speaker ID1016 (red) and ID1020 (magenta) are consistently projected to have proximity. These two points reveal a hierarchical structure of speaker embeddings, which sheds some light on the success of our methods.

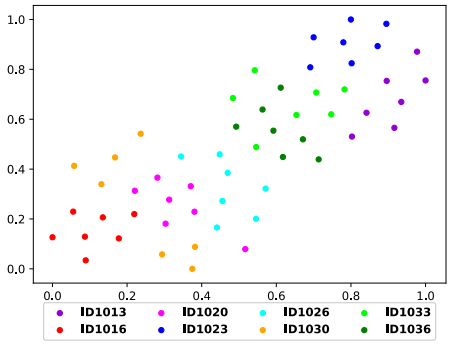
In DLKD, students and teacher networks are required to have the same embedding dimension, which limits the compression space of the student network.



(a) *x-vec+asoftmax +rhythm*



(b) *student-32*



(c) *student-32-TS*

Figure 6 T-SNE plots of 56 utterances from 8 selected speakers for different networks.

6. CONCLUSIONS

In this paper, rhythm features are introduced to reflect the distribution of phonemes and help improve the performance of speaker verification, especially intra-speaker verification. And multi-task knowledge distillation is proposed to boost the performance of the student network on both intra- and inter-speaker verifications. The embedding-level knowledge distribution directly guides the convergence of the student network. The label-level knowledge distillation transfers the posterior probabilities distribution of the incorrect outputs from the teacher network, which provides information on the similarity between speaker categories. Results show that a student can achieve a 32% relative EER reduction by using only 14.9% of parameters used in the teacher via our methods. Besides, a highly compact networks with competitive performance with the teacher network can also be obtained.

In the future, we consider further investigating the trade-off relation between the compactness and performance of student networks. Besides, the impact of angular margin loss on knowledge distillation also deserves further experiments.

ACKNOWLEDGEMENTS

Many thanks to Dr. Wei Shaojun for his valuable guidance in every stage of the writing of this paper. Thanks also to Mrs. Song for her encouragement and support.

REFERENCES

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

[2] Patrick Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal, (Report) CRIM-06/08-13, vol. 14, pp. 28–29, 2005.

[3] Najim Dehak, Patrick J Kenny, Re´da Dehak, Pierre Du- mouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2010.

[4] Shuai Wang, Zili Huang, Yanmin Qian, and Kai Yu, "Deep discriminant analysis for i-vector based robust speaker verification," in 2018 11th International Sym- posium on Chinese Spoken Language Processing (ISC- SLP). IEEE, 2018, pp. 195–199.

[5] Zili Huang, Shuai Wang, and Kai Yu, "Angular soft- max for short-duration text-independent speaker verifi- cation.," in Interspeech, 2018, pp. 3623–3627.

[6] Chunlei Zhang and Kazuhito Koishida, "End-to-end text-independent speaker verification with triplet loss on short utterances," 08 2017.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker verification," in 2018 IEEE International Con- ference on Acoustics, Speech and Signal Processing (I- CASSP), April 2018, pp. 5329–5333.

[8] EhsanVariani,XinLei,ErikMcDermott,IgnacioLopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4052–4056.

[9] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in Sixteenth annual conference of the international speech communication association, 2015.

[10] L. Yu, J. Yu, and Q. Ling, "Bltrcnn-based 3-d articulatory movement prediction: Learning articulatory synchronicity from both text and audio inputs," IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1621– 1632, July 2019.

[11] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[12] Shuai Wang, Yexin Yang, Tianzhe Wang, Yanmin Qian, and Kai Yu, "Knowledge distillation for small foot-print deep speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Sig- nal Processing (ICASSP)*. IEEE, 2019, pp. 6021–6025.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distill- ing the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Han- nemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech verification toolkit," in *IEEE 2011 workshop on automatic speech verification and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[15] Adrian Leemann, Marie-Jose´ Kolly, and Volker Dell- wo, "Speaker-individuality in suprasegmental tempo- ral features: Implications for forensic voice compari- son," *Forensic science international*, vol. 238, pp. 59– 67, 2014.

[16] Volker Dellwo, Adrian Leemann, and Marie-Jose´ Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *Thirteenth Annual Conference of the Inter- national Speech Communication Association*, 2012.

[17] Volker Dellwo and Adrian Fourcin, "Rhythmic charac- teristics of voice between and within languages," *Revue Tranel (Travaux neucha ˆtelois de linguistique)*, vol. 59, pp. 87–107, 2013.

[18] Volker Dellwo, Adrian Leemann, and Marie-Jose´ Kol- ly, "Rhythmic variability between speakers: Articula- tory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513–1528, 2015.

[19] Moez Ajili, Jean-Franc¸ois Bonastre, and Solange Rossato, "Voice comparison and rhythm: Behavioral d- ifferences between target and non-target comparisons," 2018.

[20] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," Proc. Interspeech 2018, pp. 2252–2256, 2018.

[21] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," Proc. Interspeech 2018, pp. 3573–3577, 2018.

[22] Cai,Weicheng,JinkunChen,andMingLi."Exploringtheencodinglayer and loss function in end-to-end speaker and language recognition system." arXiv preprint arXiv:1804.05160 (2018).

[23] Huang, Zili, Shuai Wang, and Kai Yu. "Angular Softmax for Short- Duration Text-independent Speaker Verification." In Interspeech, pp. 3623-3627. 2018.

[24] S.Ioffe,"Probabilisticlineardiscriminantanalysis,"inProceedingsofthe 9th European Conference on Computer Vision, ECCV 2006, ser. LNCS, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3954. Graz, Austria: Springer-Verlag Berlin, Heidelberg, may 2006, pp. 531–542.

[25] S. J. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in Proceedings of the IEEE International Conference on Computer Vision, ICCV 2007. Rio de Janeiro, Brazil: IEEE, oct 2007, pp. 1–8.

[26] Yu, Ruichi, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I. Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S. Davis. "Nisp: Pruning networks using neuron importance score propagation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9194-9203. 2018.

[27] Tai, Cheng, Tong Xiao, Yi Zhang, and Xiaogang Wang. "Convolutional neural networks with low-rank regularization." arXiv preprint arXiv:1511.06067 (2015).

[28] Ng, Raymond WM, Xuechen Liu, and Pawel Swietojanski. "Teacher-student training for text-independent speaker recognition." In 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 1044-1051. IEEE, 2018.

[29] Wang, Shuai, Yexin Yang, Tianzhe Wang, Yanmin Qian, and Kai Yu. "Knowledge distillation for small foot-print deep speaker embedding." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6021-6025. IEEE, 2019.

[30] L. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 8, pp. 2579–2605, 2008.

AUTHORS

Ruyun Li received the B.S. degree from Beihang University, Beijing, China, in 2017. She is currently pursuing the M.S. degree with the Institute of Microelectronics, Tsinghua University. Her current research interests include speech signal processing and speaker verification.

Peng Ouyang received the B.S. degree in electronic and information technology from Central South University, Changsha, Hunan, China, in 2008, and the Ph.D. degree in electronic science and technology from Tsinghua University in 2014. He held a postdoctoral position with the School of Information, Tsinghua University. He is currently the Chief Technology Officer (CTO) of TsingMicro Intelligent Technology Co. Ltd. His research interests include the embedded deep learning, neuron computing, and reconfigurable computing.

Dandan Song received the B.S. degree in automation from Harbin Engineering University, Harbin, China, in 2014, and the M.S. degree from the Institute of Microelectronics, Tsinghua University, in 2018. Her research interests include speech verification and machine learning.

Shaojun Wei was born in Beijing, China, in 1958. He received the Ph.D. degree from the Faculte Poly-technique de Mons, Belgium, in 1991. He became a Professor with the Institute of Microelectronics, Tsinghua University, in 1995. His main research interests include VLSI SoC design, EDA methodology, and communication ASIC design. He is a Senior Member of the Chinese Institute of Electronics (CIE).