# FULL-INFO TRAINING FOR DEEP SPEAKER FEATURE LEARNING

*Lantian Li, Zhiyuan Tang, Dong Wang**

Center for Speech and Language Technologies, Tsinghua University

## ABSTRACT

In recent studies, we have shown that speaker patterns can be learned from very short speech segments (e.g., 0.3 seconds) by a carefully designed convolution & time-delay deep neural network (CT-DNN) model. By enforcing the model to discriminate the speakers in the training data, frame-level speaker features can be derived from the last hidden layer. In spite of its good performance, a potential problem of the present model is that it involves a parametric classifier, i.e., the last affine layer, which may consume some discriminative knowledge, thus leading to 'information leak' for the feature learning. This paper presents a full-info training approach that discards the parametric classifier and enforces all the discriminative knowledge learned by the feature net. Our experiments on the *Fisher* database demonstrated that this new training scheme can produce more coherent features, leading to consistent and notable performance improvement on the speaker verification task.

***Index Terms***— speaker recognition, deep neural network, speaker feature learning

## 1. INTRODUCTION

Automatic speaker verification (ASV) is an important biometric authentication technology and has found a broad range of applications [1, 2]. The current ASV methods can be categorized into two groups: the statistical model approach that has gained the most popularity [3, 4, 5], and the neural model approach that emerged recently but has attracted much interest [6, 7, 8].

Perhaps the most famous statistical model is the Gaussian mixture model-universal background model (GMM-UBM) [3]. It factorizes the variance of speech signals by the UBM, and then models individual speakers conditioned on that factorization. Subsequent models design subspace structures to improve the statistical strength, including the joint factor analysis approach [4] and the i-vector model [5]. Further improvements were obtained by either discriminative models (e.g., SVM [9] and PLDA [10]) or phonetic knowledge transfer (e.g., the DNN-based i-vector method [11, 12]).

The neural model approach was also studied for many years [13, 14], however it was not as popular as the statistical model approach until recently training large-scale neural models becomes feasible. The primary success was reported by Ehsan et al. on a text-dependent task [6], where frame-level speaker features were extracted from the last hidden layer of a deep neural network (DNN), and utterance-based speaker representations ('d-vectors') were derived by averaging the frame-level features. Learning frame-level speaker features is a key merit, which paves the way to deeper understanding of speech signals. This direction, however, was not further investigated, as researchers quickly found that the relatively low performance of the d-vector approach is due to the simple back-end, i.e., the average-based utterance representation. Therefore, many researchers turn to seek for more complicated back-end models, e.g., Liu et al. [15] used DNN features to build conventional i-vector systems. Other researchers focused on the end-to-end approach that learned utterance-level representations directly, e.g., [7, 16, 17].

In spite of the reasonable success of these 'fat back-end' methods, we follow the feature learning direction originated by Ehsan et al. [6]. Our assumption is that if speaker traits are short-time identifiable and can be learned at the frame-level, many speech processing tasks will be much simpler, including ASV. Fortunately, our recent study showed that this frame-level speaker feature learning is feasible: with a short speech segment (0.3 seconds), highly representative speaker features can be learned by a convolution & time-delay DNN (CT-DNN) structure [8]. Further study showed that these speaker features are rather powerful: they can discriminate speakers by a short cough or laugh [18], and can they work well in cross-lingual scenarios [19]. We also carefully compared the feature learning approach and the end-to-end approach, and found that the feature learning approach generally works better, partly due to the more effective training scheme [20].

This paper follows the deep feature learning thread and extends our previous work in [8]. The motivation is that the present CT-DNN architecture involves a parametric classifier (i.e., the last affine layer) when training the feature learning component, or feature net. This means that part of the knowledge involved in the training data is used to learn a classifier that will be ultimately thrown away, leading to potential 'information leak'. This paper will present a full-info training approach that removes the parametric classifier so enforces all the discriminative knowledge to be learned by the feature net. Our experiments on the *Fisher* database demonstrated that this new training scheme can produce more coherent features, leading to notable and consistent performance improvement on the speaker verification task.

This work was supported by the National Natural Science Foundation of China under Grant No.61371136 / 61633013 and the National Basic Research Program (973 Program) of China under Grant No.2013CB329302. L.T. Li and Z.Y. Tang are joint first authors. Dong Wang is the corresponding author.

In the next section, we will briefly describe the CT-DNN model that we proposed for speaker feature learning, and then present the full-info training approach in Section 3. The experiments are reported in Section 4, and the paper is concluded in Section 5.

## 2. DEEP SPEAKER FEATURE LEARNING

Fig. 1 shows the CT-DNN structure that was presented in [8] and has been demonstrated in several studies [18, 19, 20]. The model consists of a convolutional (CN) component and a time-delay (TD) component. The output of the TD component is projected into a feature layer. The activations of the units of this layer, after length normalization, form frame-level speaker features. During model training, the feature layer is fully connected by an affine function to an output layer whose units correspond to the speakers in the training data, which is essentially a classifier that discriminates the target speakers in the training data on the basis of the input speech frame. The training is performed to maximize the cross entropy of the classifier output and the ground truth label. We have demonstrated that the speaker features inferred by the CT-DNN structure is highly discriminative [8], confirming our conjecture that speaker traits are largely short-time spectral patterns and can be identified at the frame level.

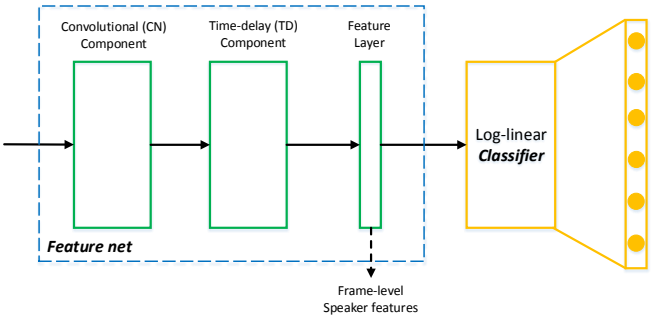


**Fig. 1**. The CT-DNN structure used for deep speaker feature learning.

## 3. FULL-INFO TRAINING

### 3.1. Method

The existing speaker feature learning models, either the vanilla structure proposed by Ehsan [6] or our CT-DNN model [8], involve two components: a *feature net* and a *classifier*, as shown in Fig. 1. The feature net produces speaker-sensitive features, and the classifier uses these features to discriminate the speakers in the training data. In the CT-DNN case, the features are produced from the last hidden layer, so the classifier is a log-linear model where the non-linear activation function is softmax. We emphasize that the feature net and the classifier are jointly trained. This is optimal if our task is to discriminate the speakers in the training set, however when the features are used in other tasks, e.g., discriminate or authenticate other speakers, the joint training will be suboptimal. This is because the classifier involves *free* parameters, so part of the discriminant information will be learned by the classifier, which, unfortunately, will be thrown away when performing identification/verification tasks on other speakers.

A possible solution is to discard the parametric classifier and using the speaker features to classify the speakers directly. Specifically, if the frame-level speaker features have been derived, each speaker $s$ in the training set can be represented by the average of all the speaker features belonging to this speaker, given by:

$$v(s;\theta) = \frac{1}{|\mathcal{E}(s)|} \sum_{x \in \mathcal{E}(s)} f(x;\theta), \tag{1}$$

where $\mathcal{E}(s)$ is the set of speech frames belonging to speaker $s$, and $f(x;\theta)$ is the speaker feature of frame $x$, produced by the feature net parameterized by $\theta$. By these speaker vectors $v(s;\theta)$, each speech frame $x$ can be classified by a simple classifier as follows:

$$p(s|f(x;\theta)) = \frac{e^{cos(f(x;\theta),v(s;\theta))}}{\sum_{s'} e^{cos(f(x;\theta),v(s';\theta))}}, \tag{2}$$

where $cos(\cdot,\cdot)$ represents cosine distance. The cross entropy between the classification output $p(s|f(x;\theta))$ and the ground truth label $s$ can be computed and used as the cost function to train the system, formulated by:

$$L(\theta) = \sum_{t} log\ p(s(t)|f(x(t);\theta)) \tag{3}$$

where $x(t)$ and $s(t)$ are the $t$-th speech frame and the corresponding ground truth label. Note that cost function involves only $\theta$ as free parameters, so all the discriminative knowledge provided by the training data is solely learned by the feature net. For this reason, we call this approach *full-info training*.

### 3.2. Implementation

Optimizing Eq. (3) is not simple, as $\theta$ appears in the speaker vectors $v(s;\theta)$. We design an iterative scheme that can perform the optimization as a usual neural net training. As shown in Fig. 2, we keep the network structure (here CT-DNN) unchanged. After each training epoch, the speaker vectors $v(s)$ are re-estimated according to Eq. (1). These speaker vectors are then used to replace the parameters of the log-linear classifier (the last affine layer), and a new epoch is started following the regular back-propagation algorithm. Note that $v(s;\theta)$ should be normalized to the same length when they are used to update the classifier, otherwise the forward computation of the last affine layer (classifier) does not equal to Eq. (2).

We experimented with various configurations to implement this iterative training, and found that allowing the parameters of the classifier to be updated *within* an epoch works slightly better than keeping them fixed. Another experience is that the a pre-trained CT-DNN model can not be used as the initial for the full-info training; a random initialization for the

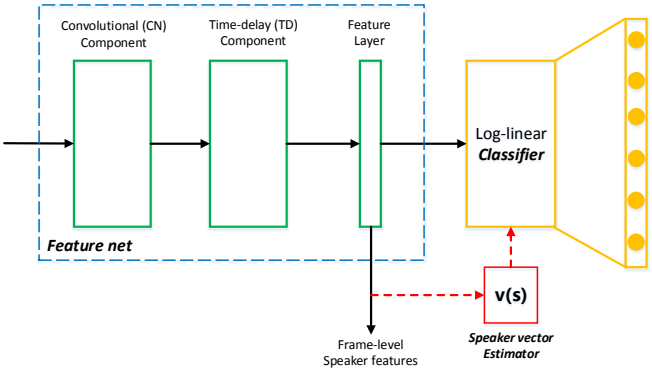


**Fig. 2**. Iterative training scheme for full-info feature learning.

feature net is required, and the training is 'warmed up' by using the speaker vectors produced by the pre-trained CT-DNN model. Once this warm-up training converges, the iterative full-info can be started.

### 3.3. Discussion

The full-info training possesses several advantages: Firstly, all the discriminant knowledge involved the training data is learned the feature net, so the training data is used more effectively; Secondly, by the full-info training, the frame-level features are encouraged to aggregate to their corresponding speaker vectors, thus more coherent; Thirdly, the distance measure used in the full-info training (cosine distance) is consistent with the measure used in the test phase of ASV. This improved consistency is important when applying speaker features to the ASV task.

We note that the full-info training has been used in some end-to-end approaches [7, 17], though we focus on learning frame-level features rather than utterance-level representations, and learn from multiple speakers rather than speaker pairs, as most of the end-to-end ASV approaches do.

## 4. EXPERIMENT

### 4.1. Database

The *Fisher* database was used in our experiments. The training set and the evaluation set are presented as follows.

- **Training set**: It consists of $2,500$ male and $2,500$ female speakers, with $95,167$ utterances randomly selected from the *Fisher* database, and each speaker has about $120$ seconds speech segments. This dataset was used for training the UBM, T-matrix, and PLDA models of the i-vector system, and the CT-DNN model of the d-vector system.

- **Evaluation set**: It consists of $500$ male and $500$ female speakers randomly selected from the *Fisher* database. There is no overlap between the speakers of the training set and the evaluation set. For each speaker, $10$ utterances are used for enrollment (about 30 seconds) and the rest for test.

We test two scenarios: a short-duration scenario and a long-duration scenario. Both scenarios involve 3 test conditions. For the short-duration scenario, the test conditions are 'S(20f)', 'S(50f)' and 'S(100f)', where the test utterances contain $20$, $50$ and $100$ frames respectively, or equivalently $0.3$, $0.6$ and $1.1$ seconds. For the long-duration scenario, the test conditions are 'L(3s)', 'L(9s)' and 'L(18s)', where the length of the test utterances is $3$, $9$ and $18$ seconds, respectively.

All the test scenarios/conditions involve pooled male- and female-dependent trials. Gender-dependent tests exhibit the same trend, so we just report the results with the pooled trials. Note that in the 'S(20f)' condition, the length of the test utterances (20 frames) is the size of the effective context window of the CT-DNN model, i.e., only one single speaker feature can be derived.

### 4.2. Settings

We build two baseline systems: an i-vector system and a d-vector system based on the CT-DNN structure. For the i-vector system, the feature involves $19$-dimensional MFCCs plus the log energy, augmented by the first and second order derivatives. The UBM consists of $2,048$ Gaussian components, and the dimensionality of the i-vector space is $400$. The entire system is trained following the Kaldi SRE08 recipe. PLDA is used in scoring.

For the d-vector system, the raw feature involves $40$-dimensional Fbanks, and a symmetric $4$-frame window is used to splice the neighboring frames. The number of output units is $5,000$, corresponding to the number of speakers in the training data. The speaker features are in $400$ dimensions, equal to the i-vectors. The utterance-level d-vectors are derived by averaging the frame-level speaker features. The Kaldi recipe to reproduce our results has been published online[1]. The scoring approach is the cosine distance on either the original $400$-dimensional d-vectors or $150$-dimensional LDA-projected vectors. Our previous experiments show that LDA can normalize the within-speaker variation, which in turn normalizes the scores of different speakers. This is important for speaker verification, as it is based on a global threshold so requires scores comparable across speakers.

### 4.3. Main results

The results in terms of equal error rate (EER%) are reported in Table 1. It can be observed that the d-vector baseline works better than the i-vector baseline in the short-term conditions, but worse in long-term conditions. This tendency is the same as in the previous studies [8, 17].

With the full-info training, we can observe that the performance of the d-vector system is improved in a consistent way. Interestingly, in the short-term test conditions, the performance with the cosine distance is not improved, but after

[1] http://project.cslt.org

LDA projection, the performance outperforms the baseline. This is indicates that the full-info training does not necessarily improve the strength of single speaker features; instead, it encourages more coherent and generalizable features. This is consistent with our discussion in Section 3.3.

**Table 1**. EER results with different models and training methods on different test conditions.

| | | EER% | | |
|---|---|---|---|---|
| Models | Scoring | S(20f) | S(50f) | S(100f) |
| i-vector | PLDA | 16.84 | 10.41 | 6.54 |
| d-vector | Cosine | 7.89 | 6.38 | 4.55 |
| | LDA | 8.15 | 5.05 | 3.38 |
| d-vector | Cosine | 9.48 | 7.45 | 4.74 |
| + Full-info Training | LDA | **7.53** | **4.36** | **2.85** |
| Models | Scoring | L(3s) | L(9s) | L(18s) |
| i-vector | PLDA | 3.52 | **1.20** | **0.89** |
| d-vector | Cosine | 3.85 | 2.90 | 2.69 |
| | LDA | 2.58 | 1.95 | 1.79 |
| d-vector | Cosine | 3.95 | 2.48 | 2.23 |
| + Full-info Training | LDA | **2.14** | 1.64 | 1.54 |

### 4.4. Analysis

#### *4.4.1. Training process*

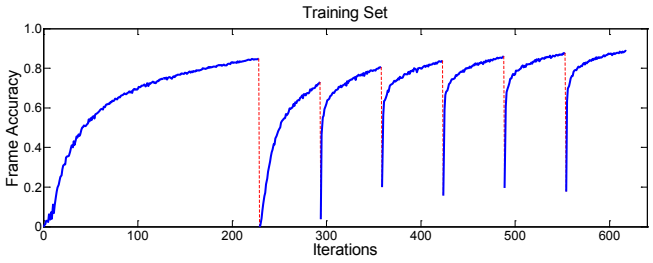


**Fig. 3**. The change of frame accuracy on the training set during the iterative full-info training.

Fig. 3 presents the change of the training-set frame accuracy during the iterative training process(the trend on the validation set is the same). It can be observed that the accuracy is increased within each epoch, and after each epoch, the initial accuracy starts from a higher value than the previous epoch. This indicates an increased coherence between the frame-level speaker features and the speaker vectors. Note that the big gap between the accuracies of the start and end stage of an epoch is due to the adaptable last affine layer.

The EER results on the 6 test conditions during the iterative training process are shown in Fig. 4. We can observe a consistent and notable EER reduction on all the test conditions.

#### *4.4.2. Visualization*

T-SNE [21] is used to visualize the speaker features in the 2-dimensional space. In Fig. 5, we choose several utterances

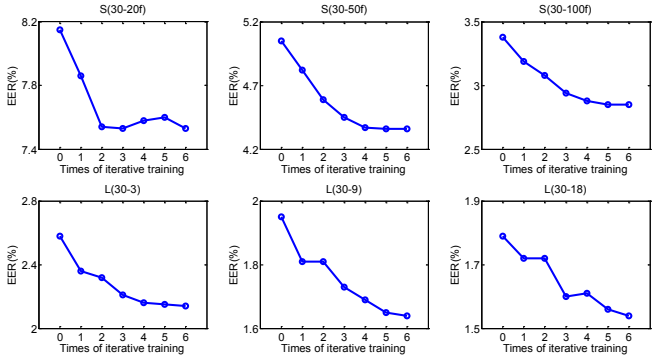


**Fig. 4**. EER(%) on the 6 test conditions with iterative full-info training.

by 20 different speakers, and draw the speaker features produced with and without the full-info training. It can be observed that the speaker features are highly discriminative, no matter whether the full-info training is applied. However, the full-info training produces more coherent features. Paying attention to the circled features, we observe that there are two speakers (red and cyan) whose features are located in separated areas in the left picture while aggregate together in the right picture. This clearly demonstrates that the full-info training encourages more coherent features.

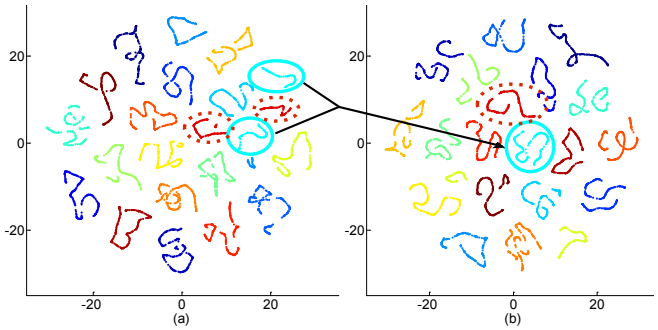


**Fig. 5**. Deep speaker features are *excerpted from a single utterance* and plotted by t-SNE, with each color representing a speaker. where (a) shows features produced by the original CT-DNN model, and (b) shows features produced by the CT-DNN model trained with full-info training.

## 5. CONCLUSIONS

This paper proposed a full-info training approach that enforces all the speaker discrimination knowledge provided by the training data being learned by the feature net, thus avoiding the 'information leak' caused by the parametric classifier involved in the conventional learning structure. We tested the method on the speaker verification task with the Fisher database, and found that it delivered consistent and notable performance improvement. Methods that encourage more coherent features are under investigation.

## 6. REFERENCES


[1] Homayoon Beigi, *Fundamentals of speaker recognition*, Springer Science & Business Media, 2011.

[2] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

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

[4] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[5] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, 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, 2011.

[6] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[7] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.

[8] Lantian Li, Yixiang Chen, Ying Shi, Zhiyuan Tang, and Dong Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[9] William M Campbell, Douglas E Sturim, and Douglas A Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.

[10] Sergey Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision–ECCV 2006*, pp. 531–542, 2006.

[11] Patrick Kenny, Vishwa Gupta, Themos Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014, pp. 293–298.

[12] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[13] Kevin R Farrell and Richard J Mammone, "Speaker recognition using neural tree networks," in *Advances in Neural Information Processing Systems*, 1994, pp. 1035–1042.

[14] Fazal Mueen, Ayaz Ahmed, A Gaba, et al., "Speaker recognition using artificial neural networks," in *Students Conference, 2002. ISCON'02. Proceedings. IEEE*. IEEE, 2002, vol. 1, pp. 99–102.

[15] Yuan Liu, Yanmin Qian, Nanxin Chen, Tianfan Fu, Ya Zhang, and Kai Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

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

[17] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev 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.

[18] Miao Zhang, Yixiang Chen, Lantian Li, and Dong Wang, "Speaker recognition with cough, laugh and" wei"," *arXiv preprint arXiv:1706.07860*, 2017.

[19] Lantian Li, Dong Wang, Askar Rozi, and Thomas Fang Zheng, "Cross-lingual speaker verification with deep feature learning," *arXiv preprint arXiv:1706.07861*, 2017.

[20] Dong Wang, Lantian Li, Zhiyuan Tang, and Thomas Fang Zheng, "Deep speaker verification: Do we need end to end?," *arXiv preprint arXiv:1706.07859*, 2017.

[21] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.