End-to-End Residual CNN with L-GM Loss Speaker Verification System

Xuan Shi, Mengyao Zhu, Xingjian Du
School of Communication and Information Engineering, Shanghai University
Shanghai, China
shixuan1996@i.shu.edu.cn, diggerdu@shu.edu.cn, zhumengyao@shu.edu.cn

***Abstract*—We propose an end-to-end speaker verification system based on the neural network and trained by a loss function with less computational complexity. The end-to-end speaker verification system consists of a ResNet architecture to extract features from utterance, then mean pool to produces utterance-level speaker embeddings, and train using the large-margin Gaussian Mixture loss function. Influenced by the large-margin and likelihood regularization, large-margin Gaussian Mixture loss function benefits the speaker verification performance. Experimental results demonstrate that the Residual CNN with large-margin Gaussian Mixture loss outperforms DNN-based i-vector baseline by nearly 10% reduction in equal error rate.**

***Index Terms*—Speaker Verification, End-to-End Training, Large-Margin Gaussian Mixture Loss**

I. Introduction

Speaker verification (SV) aims to determine whether the utterance comes from the claimed identity or not. Verification algorithm may require the speaker to utter a specific phrase (text-dependent) or be agnostic to the audio transcript (text-independent). In text-independent SV, no prior constraints are considered for the spoken phrases by the speaker, which makes it challenging compared to text-dependent scenario.

The conventional speaker verification approach entails using i-vectors [1] and probabilistic linear discriminant analysis (PLDA) [2]. As a supervised learning method, i-vector requires sufficient statistics which are computed from a Gaussian Mixture Model-Universal Background Model (GMM-UBM), followed by a PLDA model to produce verification scores [1]. Recently, inspired by using deep neural network in Automatic Speech Recognition(ASR) [3], other research efforts have been conducted on the application of DNN in speaker verification. DNN was used to extract abundant statistics and convert them from high dimension to a low-dimension vector, followed by a PLDA or SVM model trained to provide a classification score.

Recently, [4] introduced an end-to-end system trained to discriminate between same-speaker and different-speaker utterance pairs. First, a deep neural network is used to extract frame-level features from utterances. Then, pooling and length normalization layers generate utterance-level speaker embeddings. The model is trained using triplet loss [5], which minimizes the distance between embedding pairs from the same speaker and maximizes the distance between pairs from different speakers. Based on the deep Residual CNN (ResNet) [6] and triplet loss [5], it outperformed an i-vector baseline. Nevertheless, triplet loss is still not effective enough because the cosine distance among triplet features are added as additional loss at each time. It inevitably results in slow convergence and instability. By carefully selecting the image triplets, the problem may be partially alleviated. But it significantly increases the computational complexity and the training procedure becomes inconvenient [7].

To circumvent the drawbacks of triplet loss, [7] introduced the center loss by minimizing the Euclidean distance between the features and the corresponding class centroids. However, Euclidean distance results in the inconsistency of distance measurements in the feature space. The large-margin Gaussian Mixture (L-GM) loss [8] was proposed to solve the drawbacks of center loss [7]. L-GM loss adopted the Mahalanobis distance to measure the distance between extracted features and the feature centroid of ground truth class, which contributed L-GM loss has a better performance in inter-class dispension and intra-class compactness.

In this paper, we extend the end-to-end speaker embedding systems proposed in [4], but replace the triplet loss with large-margin Gaussian Mixture (L-GM) loss [8]. We use L-GM loss under the assumption that the embeddings of speaker utterances follow a Gaussian mixture distribution.

Finally, we evaluate our speaker verification system on dataset VoxCeleb [9] to investigate the performance. In the best case (α = 1), ResNet + L-GM loss performs better than DNN-based i-vector system, which reduces the verification equal error rate by nearly 10%.

II. Related work

Traditionally, researchers tend to create Gaussian mixture model(GMM)-based speaker verification systems. The most fundamental GMM-based speaker verification methods include the classical maximum a posteriori (MAP) adaptation of universal background model parameters (GMM-UBM) [10] [11] [12] and support vector machine (SVM) modeling of GMM super-vectors (GMM-SVM) [13].

I-vector system was proposed in [1], which is also a state-of-art and high-effective system to verify speaker's identity. I-vector-based speaker verification models perform classification using cosine similarity between i-vectors or more advanced techniques such as PLDA, heavy-tailed PLDA , and Gauss-PLDA.

Recently, the solution of speaker verification task has increasingly been considered from the perspective of deep

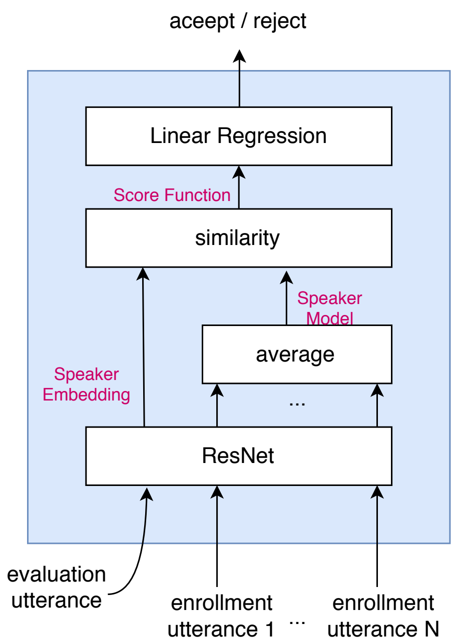


Fig. 1: Full architecture

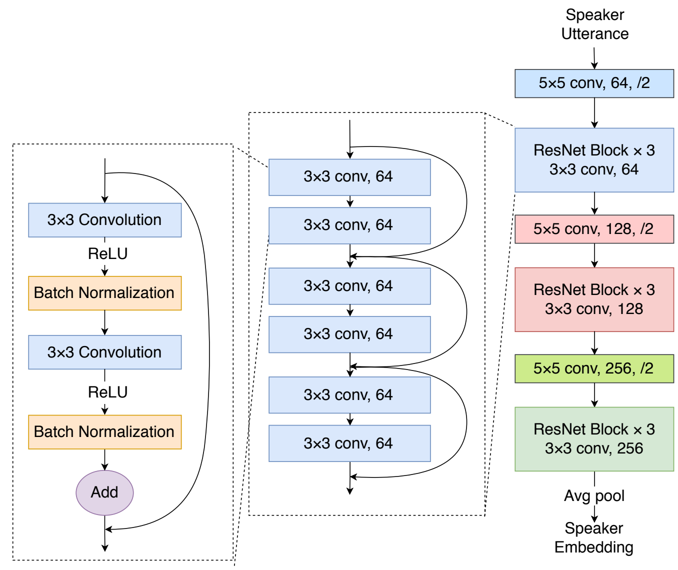


Fig. 2: ResNet architecture

learning approaches. There are several models replace the components of traditional SV systems with DNN or other neural network architectures. For example, [14] utilized a Deep Neural Network/Hidden Markov Model Automatic Speech Recognition (DNN/ HMM ASR) system to extract content-related posterior probabilities from utterances. [14] proposed an transfer learning method based on Bayesian joint probability to help find a better optimal solution of PLDA parameters for the target domain.

A growing number of papers presented end-to-end neural networks for speaker verification. [15] utilized the ResNet with spectrograms as an input features in the text-dependent speaker verification task and similar architecture was used in [4] for text-independent speaker verification study.

III. Speaker Verification System

A. Overview

The proposed architecture bases a Residual CNN(ResNet) that extracts statistics from utterances in training set and maps them to speaker embeddings. The objective function operates one utterance at one time to compact intra-class variations and separable inter-class differences as much as possible. After the training stage is complexed, the parameters of ResNet is fixed. For enrollment, the speaker model is used to generate embeddings from each speakers. Finally, during the evaluation stage, scoring function provides the similarity of utterances between claimed speaker and input. Fig.1 shows the overview of our system.

B. Residual CNN

Deep neural networks perform better than shallow networks in extracting features, but it is not easy to train them. Compared with deep neural networks, ResNet [6] are easier to optimize and can gain accuracy from considerably increased depth. ResNet is comprised by ResBlock, which is defined as:

$$h = F(x, W_i) + x \tag{1}$$

The x and h denote the input and the output of the ResNet block. F is the stacked nonlinear layer's mapping function. And W_i means the i-th weight of the mapping function F. The formulation of $F(x, W_i) + x$ can be realized by feedforward neural networks with "shortcut connections".

Based on the original ResNet architecture, we vary the size of filter and stride for each ResBlock shown in Fig.2. Thus, each block owns an identical structure and the shortcut is the identity mapping of x. Three ResBlocks are stacked in an architecture and the number of channels double. When the number of channels increases, we use a single convolutional layer with filter size of 5×5 and a stride of 2×2.

C. Speaker Embedding

Although an embedding can be extracted from an utterance as long as well in theory, it is constrained by memory practically. We adopt a expedient method by extracting embeddings from 20 second chunks so that memory could be exploited and features are sufficient for our model to make decisions. A single embedding is generated from the entire utterance if it is shorter than 20 seconds. Enrollment embeddings are extracted from one or more utterances, and averaged to create a speaker-level representation. Enroll and evaluate utterances are scored by the distance metric used in the objective function.

D. Large-Margin GM Loss

In the deep speaker system [4], the triplet loss function [5] was utilized as loss function in speaker verification task, which indicates the similarity between speaker verification task and face verification task. In the realm of face verification, L-GM loss [8] is a more efficient and robust loss function, so it is reasonable to investigate the performance of L-GM loss function in speaker verification task. However, only when the extracted deep feature x on the training set follows a Gaussian mixture distribution will the L-GM could well perform [8]. To fit this precondition, we assume that voiceprint based on the physical configuration of a speaker's mouth and throat keeps steady with slight fluctuation in a certain time, which corresponds to the intuition of Gaussian distribution. We elaborate the large-margin Gaussian Mixture loss function mathematically according to [8].

Different from the triplet loss [5], we hereby assume that the embedding x on the training set follows a Gaussian mixture distribution expressed in Eq.2. μ_k and Σ_k are the mean and covariance of speaker k in the embedding space; and $p(k)$ is the prior probability of speaker k.

$$p(x) = \sum_{k=1}^{K} \mathcal{N}(x; \mu_k, \Sigma_k) p(k) \tag{2}$$

Under such an assumption, the conditional probability distribution of an embedding x_i given its class label $z_i \in [1, K]$ can be expressed in Eq. 3. Consequently, the corresponding posterior probability distribution can be expressed in Eq. 4.

$$p(x_i|z_i) = \mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}) \tag{3}$$

$$p(z_i|x_i) = \frac{\mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}) p(z_i)}{\sum_{k=1}^{K} \mathcal{N}(x_i; \mu_k, \Sigma_k) p(k)} \tag{4}$$

As such, a *classification loss* $\mathcal{L}_{cls}$ can be computed as the cross-entropy between the posterior probability distribution and the one-hot class label as is shown in Eq. 5, in which the indicator function $\mathbb{1}()$ equals 1 if z_i equals k; or 0 otherwise.

$$\begin{aligned} \mathcal{L}_{cls} &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}(z_i = k) \log p(k|x_i) \\ &= -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}) p(z_i)}{\sum_{k=1}^{K} \mathcal{N}(x_i; \mu_k, \Sigma_k) p(k)} \end{aligned} \tag{5}$$

However, extracted feature x_i may be far away from the corresponding class centroid μ_{z_i} while still being correctly classified as long as it is relatively closer to μ_{z_i} than to the feature means of the other classes. To solve this problem, it is necessary to add a *likelihood regularization* term, defined as the sum of negative log likelihood in Eq.6, measuring to what extent the training samples fit the assumed distribution:

$$\mathcal{L}_{lkd} = -\sum_{i=1}^{N} \log \mathcal{N}(x_i; \mu_{z_i}, \Sigma_{z_i}) \tag{6}$$

Finally the proposed GM loss $\mathcal{L}_{GM}$ is defined in Eq. 7, in which λ is a non-negative weighting coefficient.

$$\mathcal{L}_{GM} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{lkd} \tag{7}$$

By definition, for the training feature space, the classification loss $\mathcal{L}_{cls}$ is mainly related to its discriminative capability while the likelihood regularization $\mathcal{L}_{lkd}$ is related to its probabilistic distribution. Under the GM distribution assumption, $\mathcal{L}_{cls}$ and $\mathcal{L}_{lkd}$ share all the parameters.

To optimize the generalization of loss function, large classification margin is applied in the training process. Denote x_i's contribution to the classification loss to be $\mathcal{L}_{cls,i}$, of which an expansion form is in Eq. 8 and Eq. 9.

$$\mathcal{L}_{cls,i} = -\log \frac{p(z_i)|\Sigma_{z_i}|^{-\frac{1}{2}} e^{-d_{z_i}}}{\sum_k p(k)|\Sigma_k|^{-\frac{1}{2}} e^{-d_k}} \tag{8}$$

$$d_k = (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)/2 \tag{9}$$

The d_k denotes the squared Mahalanobis distance which is obviously non-negative. Then, a classification margin $m \geq 0$ is added to d_k so that $\mathcal{L}_{cls,i}$ gets large margin.

$$\mathcal{L}_{cls,i}^{m} = -\log \frac{p(z_i)|\Sigma_{z_i}|^{-\frac{1}{2}} e^{-d_{z_i} - m}}{\sum_k p(k)|\Sigma_k|^{-\frac{1}{2}} e^{-d_k - \mathbb{1}(k=z_i)m}} \tag{10}$$

The x_i is classified to the class z_i if and only if Eq. 11 holds, indicating that x_i should be closer to the feature mean of class z_i than to that of the other classes by at least m.

$$e^{-d_{z_i} - m} > e^{-d_k} \iff d_k - d_{z_i} > m \quad ,\forall k \neq z_i \tag{11}$$

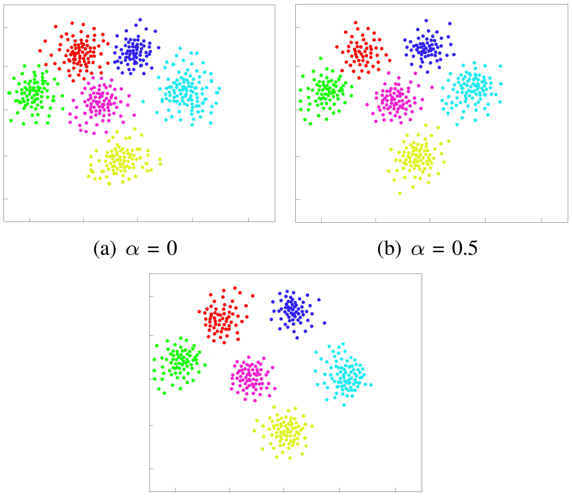
It is in a dilemma to fix the value of margin m properly. On one hand, a large margin could significantly force the features of different classes apart and pull the features of same class to their feature mean of class. On the other hand, a larger margin may cause a difficulty of optimizing when number of classes gets lager. Therefore, introducing an adaptive scheme for designing the margin is quite necessary. An adaptive scheme is proposed by setting the value of m to be proportional to each sample's distance to its corresponding class feature mean, i.e., $m = \alpha d_{z_i}$, in which α is a non-negative parameter controlling the size of the expected margin between two classes on the training set [8]. Fig. 3 shows a schematic interpretation of α.

IV. EXPERIMENT

In this sector we present out experiment setup, as well as details related to the ResNet architecture, L-GM loss, and neural network training.

A. Dataset

In this paper, we run our speaker verification system on dataset VoxCeleb [9] to investigate the performance. VoxCeleb contains over 100,000 utterances for 1,251 celebrities, extracted from videos uploaded to YouTube. The dataset is gender balanced, with 55% of the speakers male. The speakers span a wide range of different ethnicities, accents, professions and ages.



(a) α = 0 (b) α = 0.5

(c) α = 1

Fig. 3: The schematic interpretation of embeddings distribution with different α

B. Experimental Setup

For verification, all Person of Internet(POIs) whose name starts with an 'E' are reserved for testing, since this gives a good balance of male and female speakers. These POIs are not used for training the network, and are only used at test time. The statistics are given in Table I.

Set	# POIs	# Vid. / POI	# Utterances
Dev	1,211	18.0	140,664
Test	40	17.4	4,715

TABLE I: Development and test set statistics for verification.

Two key performance metrics are used to evaluate system performance for the verification task. The metrics are similar to those used by existing datasets and challenges, such as SITW [16]. The primary metric is based on the cost function C_{det}

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \quad (12)$$

where we assume a prior target probability P_{tar} of 0.01 and equal weights of 1.0 between misses C_{miss} and false alarms C_{fa}. The primary metric, C_{det}^{min}, is the minimum value of C_{det} for the range of thresholds. The alternative performance measure used here is the Equal Error Rate (EER) which is the rate at which both acceptance and rejection errors are equal. This measure is commonly used for identity verification systems [9].

C. Baselines

1) DNN-based i-vector: In this paper, we investigate the effect of L-GM loss in the speaker verification task which is mainly different from the deep speaker system [4]. Lacking of detailed but vital parameters in [4], we failed to set the deep speaker system as our baseline. Thus, we attempt to compare with the DNN i-vector, the baseline of deep speaker system, for the performance in SV tasks.

The baseline DNN i-vector model is built based on [17]. A seven-layer DNN with 600 input nodes, 1200 nodes in each hidden layer and 3450 output nodes was trained with cross entropy using the alignments from the HMM-GMM. The input layer of the DNN is composed of 15 frames (7 frames on each side of the frame for which predictions are made) where each frame corresponds to 40 log Mel-filterbank coefficients. The DNN is used to provide the posterior probability in the proposed framework for the 3450 senones defined by a decision tree.

D. Experiments

In this section, we conduct the speaker verification experiments to investigate the performances of large-margin Gaussian Mixture loss and the influence of α. In our experiments, we empirically set α to 1.0, 0.3, 0.1, 0.01, and 0 for our speaker verification tasks. We also set the regularization λ to a small value so that the $\mathcal{L}_{lkd}$ can truly benefit the training process and improve the training accuracy. All experiments are carried out using the tensorflow framework [18].

E. Results

The Tables II shows the performance of DNN i-vector system and end-to-end speaker verification system. It is apparent that ResNet system has reached the performance of DNN i-vector system(α = 1). Also, there is an obvious improvement of C_{det}^{min} after altering α from 0 to 1, which confirms that increasing the margin parameter α benefits the speaker verification performance [8].

Metrics	C_{det}^{min}	EER (%)
DNN I-vectors	0.73	8.8
ResNet + L-GM Loss(α = 0)	0.81	15.3
ResNet + L-GM Loss(α = 0.01)	0.78	12.9
ResNet + L-GM Loss(α = 0.1)	0.75	10.5
ResNet + L-GM Loss(α = 0.3)	0.74	9.8
ResNet + L-GM Loss(α = 1)	0.71	7.9

TABLE II: Results for verification on `VoxCeleb`

V. Conclusion

The insight of our work does not claim too much effort to understand: a decent and practical speaker verification system relies on an robust feature extractor to retrieve the intrinsic and distinctive embedding of speaker's voice and a well-designed metric to evaluation the sparsity of embedding space. In this work, we investigate the efficiency of large-margin Gaussian Mixture loss for model training and the application of ResNet for feature extraction. The combination of ResNet and large-margin Gaussian Mixture loss function show a promised performance in our dataset and surpass the baseline system with a considered margin in terms of equal error rate.

Acknowledgment

This work was supported by the key support Projects of Shanghai Science and Technology Committee (16010500100).

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] S. Cumani, O. Plchot, and P. Laface, “Probabilistic linear discriminant analysis of i-vector posterior distributions,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*. IEEE, 2013, pp. 7644–7648. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6619549

[3] “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[4] 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,” *CoRR*, vol. abs/1705.02304, 2017. [Online]. Available: http://arxiv.org/abs/1705.02304

[5] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 815–823. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7293313

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.

[7] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9911. Springer, 2016, pp. 499–515. [Online]. Available: https://doi.org/10.1007/978-3-319-46478-7

[8] W. Wan, Y. Zhong, T. Li, and J. Chen, “Rethinking feature distribution for loss functions in image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 2616–2620. [Online]. Available: https://doi.org/10.21437/Interspeech.2017

[10] B. G. B. Fauve, D. Matrouf, N. Scheffer, J. Bonastre, and J. S. D. Mason, “State-of-the-art performance in text-independent speaker verification through open-source software,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 15, no. 7, pp. 1960–1968, 2007.

[11] S. Yin, R. C. Rose, and P. Kenny, “A joint factor analysis approach to progressive model adaptation in text-independent speaker verification,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 15, no. 7, pp. 1999–2010, 2007.

[12] T. Hasan and J. H. L. Hansen, “A study on universal background model training in speaker verification,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 1890–1899, 2011.

[13] W. M. Campbell, J. P. Campbell, T. P. Gleason, D. A. Reynolds, and W. Shen, “Speaker verification using support vector machines and high-level features,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 15, no. 7, pp. 2085–2094, 2007.

[14] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5115–5119. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7465907

[15] E. Malykh, S. Novoselov, and O. Kudashev, “On residual CNN in text-dependent speaker verification task,” in *Speech and Computer - 19th International Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017, Proceedings*, ser. Lecture Notes in Computer Science, A. Karpov, R. Potapova, and I. Mporas, Eds., vol. 10458. Springer, 2017, pp. 593–601. [Online]. Available: https://doi.org/10.1007/978-3-319-66429-3

[16] M. McLaren, L. Ferrer, D. Castán, and A. Lawson, “The speakers in the wild (SITW) speaker recognition database,” in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 818–822. [Online]. Available: https://doi.org/10.21437/Interspeech.2016

[17] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 2014, pp. 1695–1699. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6844297

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org.