2018 IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin).

Discriminative Deep Audio Feature Embedding for Speaker Recognition in the Wild

Simone Bianco, Elia Cereda, Paolo Napoletano
Department of Informatics, Systems and Communication
University of Milano-Bicocca
viale Sarca, 336, 20126, Milano, Italy
Email: {bianco, napoletano}@disco.unimib.it, e.cereda7@campus.unimib.it

***Abstract*—In this paper we face the problem of speaker recognition in the wild. We tackle the speaker identification and verification problems with the use of Deep Convolutional Neural Networks (CNN). We propose the modification of two Residual CNN architectures (ResNet) in order to be used with the spectrograms of the audio data as input images. The proposed architectures, trained with a contrastive-center loss, have been tested on the VoxCeleb and SIWIS datasets on both the speaker identification and verification tasks. The experimental results show the effectiveness of the proposed solution with respect to the state of the art. The proposed network shows to be robust in unconstrained conditions and, more important, it shows to be quite robust in a multilingual scenario.**

***Index Terms*—Speaker Identification, Speaker Verification, Convolutional Neural Networks.**

I. Introduction

Speaker recognition includes both the tasks of speaker identification and verification [1]. Speaker identification consists in the identification of the speaker among a closed set of speakers. As such, it is a n-ary classification problem, where n is the size of the set of speakers. On the other hand, speaker verification is a binary classification problem, and consists in the verification of the claimed identity. Both speaker identification and verification are two challenging topics, especially under unconstrained conditions. Possible applications range from voice-based authentication systems to speaker diarization, from forensic trials to personalized AI assistants. Verification and identification algorithms in the state of the art belong to two classes: text-dependent, i.e. they require the speaker to pronounce a specific sentence (or sequence of numbers), or text-independent, i.e. they are agnostic to the sentence pronounced by the speaker.

The speaker identification problem has been faced for a long time with the use of handcrafted features obtained through the use of Gaussian of Mixture Models (GMMs) [2], [3] combined with Joint Factor Analysis (JFA) [4] and i-vectors [5]. These methods allow to define a new low-dimensional space that models both speaker and channel variabilities. The scoring method is then based on cosine similarity in the new space. All these models rely on low dimensional representations of the audio signal that are rather sensitive to noise (especially present in real world applications) and, more importantly, these representations lack in speaker-discriminating features, such as pitch information.

Deep Convolutional Neural Networks (CNNs) have been successfully applied to speech recognition using higher dimensional inputs [6], [7] and to speaker identification [8]. CNN-based approaches have replaced handcrafted-features-based methods mostly because of their ability to be robust against the noise of real world data. The CNNs trend is also supported by the release of a very large public dataset for speaker identification: the VoxCeleb [9] dataset, that is a text-independent speaker dataset containing hundreds of thousands of "real world" utterances for over 1,000 celebrities.

The paper by Li et al. [10] proposed a neural speaker embedding system that maps utterances to a hypersphere where speaker similarity is measured by cosine similarity.

Nagrani et al. [9] proposed a modified version of a CNN architecture named VGG-M (Visual Geometry Group Medium), that was originally proposed by Chatfield et al. [11] for image classification. The VGG-M is used by Nagrani et al. for the identification task, while for verification the same architecture is used to build a Siamese network that is trained using a contrastive loss that learns an embedding of the speaker.

Here we propose a CNN architecture based on Residual Networks (ResNet) to face with both the speaker identification and verification task. Residual architectures, that have been originally introduced for image classification by He et al. [12], are based on the idea that each layer of the network learns residual functions with reference to the layer inputs instead of learning unreferenced functions. He et al. [12] demonstrated that such architectures are easier to optimize, obtain better accuracy and are less prone to overfitting even at considerably higher depths. These architectures demonstrated to perform much better than previous architectures, including VGG-based ones, on the image classification task [12].

The main idea of this paper is to take advantage of the contrastive-center loss function to train the ResNet in order to obtain highly discriminative feature embeddings. The trained network is used end-to-end for the speaker identification task while the network features are used jointly with a cosine similarity scoring function for speaker verification.

II. Proposed approach

Following the work from [9] on the VGG architecture, we adapted two networks of the ResNet [12] family, originally introduced for image classification, to the problem of speaker

978-1-5386-6095-9/18/$31.00 ©2018 IEEE

TABLE I
SIDE-BY-SIDE ARCHITECTURE COMPARISON OF THE MODIFIED VGG-M [9] AND THE PROPOSED MODIFIED RESNET-18 AND RESNET-34

Layer name	**VoxCeleb VGG-M** [9]	**Proposed ResNet-18**	**Proposed ResNet-34**
conv1	7×7, 96, stride 2	7×7, 64, stride 2	
mpool1	3×3 max pool, stride 2		
conv2	5×5, 256, stride 2	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$
mpool2	3×3 max pool, stride 2	–	
conv3	3×3, 256, stride 1	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$
conv4	3×3, 256, stride 1	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 6$
conv5	3×3, 256, stride 1	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$
mpool5	5×3 max pool, stride 3×2	–	
fc6	9×1, 4096, stride 1	16×1, 4096, stride 1	
apool6	1×n average pool, stride 1		
fc7	1×1, 1024, stride 1		
fc8	1×1, 1211, stride 1		

identification. The proposed architecture is trained using a linear combination of two losses: contrastive-center loss and softmax cross entropy. This allows us to build a network trained for speaker identification that embeds discriminative features. Furthermore, these features can be directly applied for speaker verification using cosine similarity without adding complexity to the training process.

A. Input Features

We apply the same pre-processing proposed by [9] to the audio data. Utterances are converted to single-channel 16-bit PCM streams and resampled at 16kHz. Spectrograms are then generated applying the Short Time Fourier Transform (STFT) with a Hamming window of width 25ms, a step of 10ms and a 1024-point Fast Fourier Transform. Mean and variance normalization is applied to every frequency bin of the spectrum. For a 3-second sample, the output is a complex matrix of size 513×300. The magnitude of the resulting spectrogram is then used as input to the CNN across all of our experiments. Figure 1 shows the example of one spectrogram generated by our input pipeline.

B. Network Architecture

The original VGG and ResNet networks were designed to work on images of a single predefined size. Spectrograms, on the other hand, vary along the time dimension depending on the length of the input recording. While a traditional approach to deal with variable-length utterances has been to split them into fixed segments and average the final predictions, the authors of [9] obtained the best results by aggregating features midway through the neural network with an average pooling layer having support $1 \times n$, where n is computed from the length of the input speech segment.

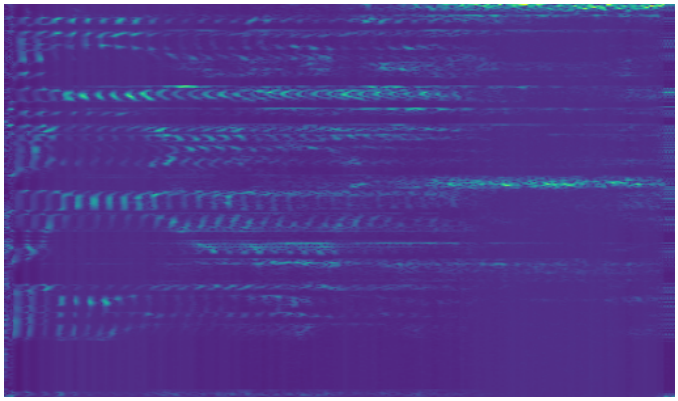


Fig. 1. Visualization of the mean- and variance-normalized magnitude spectrogram of a 3-second sample from the `SIWIS` dataset. Time increases top to bottom, frequency increases left to right (the image is transposed for visualization).

We replicate this strategy on ResNet-18 and ResNet-34, replacing the single fully connected layer in the original architecture with layers from *fc6* to *fc8*, inspired by the VoxCeleb VGG. Table I shows the complete structure of the three networks.

For speaker identification, softmax is applied to the output of the last layer *fc8* to produce a probability distribution over the target speakers in the dataset.

For verification, we process the utterances with the same network trained for identification by extracting and l^2-normalizing the activations of the layer *fc7*. These features are embedded in a 1024-dimensional feature vector space where

cosine similarity is used to compare the speaker identities.

C. Contrastive-Center Loss

An effective technique to train CNNs that learn highly discriminative feature embeddings is to use an auxiliary loss function that directly optimizes similarity among the features of a target class. Some functions that have been applied for this purpose are contrastive loss [13], triplet loss [14], and center loss [15]. All three are meant to produce features with better intra-class compactness and inter-class separability. The contrastive loss and the triplet loss are effective at this task, but at the cost of complicating training: they operate on pairs or triplets of samples which must be chosen very carefully to ensure good metric learning. Furthermore, the number of pairs scales quadratically with the number of samples, making it impractical to select all possible combinations. The center loss is a novel loss function that addresses this issue by operating sample-wise. It learns a center for each class and penalizes samples with a high distance from their respective class center, improving intra-class compactness. It does not however optimize for inter-class separability.

The contrastive-center loss [16] is an evolution of the center loss that simultaneously considers both properties by penalizing the ratio between the distances of each sample from its corresponding class center and the sum of the distances from all other centers, as shown in Eq. 1:

$$L_{ct-c} = \frac{1}{2}\sum_{i=1}^{m} \frac{\|x_i - c_{y_i}\|_2^2}{\left(\sum_{j=1, j\neq y_i}^{k} \|x_i - c_{y_i}\|_2^2\right) + \delta} \qquad (1)$$

where m is the number of training samples in a mini-batch, $x_i \in \mathbb{R}^d$ is the i-th training sample and y_i its corresponding class, $c_{y_i} \in \mathbb{R}^d$ denotes the y_i-th class center, d is the embedding size, k is the number of classes and δ is a constant added to prevent the denominator from becoming 0.

The CNN is then trained with the joint supervision of the softmax loss L_s and the contrastive-center loss L_{ct-c}, using a scalar λ to balance their contribution. The total loss of the network is then:

$$L = L_s + \lambda L_{ct-c} \qquad (2)$$

In our experiments we set $\delta = 1$ and $\lambda = 0.1$, the values proposed by [16].

This formulation allows the class centers c_{y_i} to be learned alongside the other network parameters, without modifying the training process.

III. Experiments

A. Datasets

In the experiments we used two different datasets: `VoxCeleb` [9] for training and evaluating performance in real-world conditions for both speaker identification and verification; `SIWIS` [17] for evaluating speaker verification in a cross-language scenario.

`VoxCeleb` is a collection of over 100,000 utterances from 1,251 celebrities, extracted from videos uploaded to YouTube. The speakers span a wide range of different ethnicities, accents, professions and ages, although being heavily skewed toward American English. Crucially, the samples of their voices included in the dataset are all degraded by real world noise, in the form of background chatter, laughter, overlapping speech or recording equipment. This is fundamental for training models that generalize successfully to the unconstrained conditions found "in the wild".

`SIWIS` is composed of speech recordings from 36 bilingual and trilingual speakers of English, French, German and Italian. The speakers were evaluated in each language by at least 3 native judges expert in linguistics and deemed effectively accentless. The reading material for every speaker-language pair consisted of a set of 171 prompts, corresponding to a recording session of approximately 20 minutes.

TABLE II
Speaker Identification results on `VoxCeleb` dataset in terms of accuracy

Accuracy	**top-1** (%)	**top-5** (%)
I-vectors + SVM [5], [9]	49.0	56.6
I-vectors + PLDA + SVM [5], [9]	60.8	75.6
VoxCeleb CNN [9]	80.5	92.1
VoxCeleb CNN (reimplemented)	81.1	91.5
VoxCeleb CNN + contrastive-center loss	82.7	92.8
ResNet-18 + contrastive-center loss	84.3	93.4
ResNet-34 + contrastive-center loss	**85.2**	**93.9**

B. Speaker Identification

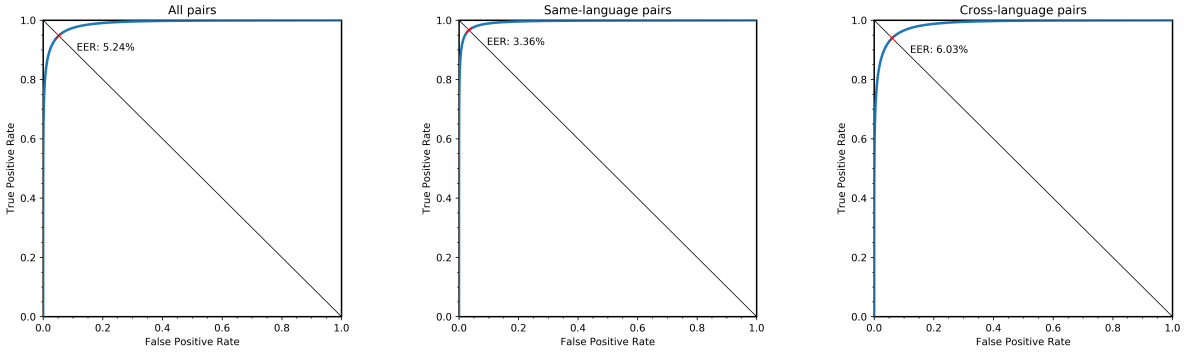
We evaluate the performance of speaker identification in terms of top-1 and top-5 accuracy on the `VoxCeleb` dataset. The results are reported in Table II. It is possible to see that the use of constrastive-center loss is able to improve the top-1 accuracy of the VoxCeleb CNN architecture [9] by 2.2%. The best top-1 accuracy is obtained by the proposed ResNet-34 architecture with contrastive-center loss, with an improvement of 4.7% with respect to the state of the art.

C. Speaker Verification

We evaluate the performance of speaker verification using the Equal Error Rate (EER) metric [1], a measure commonly used for identity verification systems. It is defined as the point on the Receiver Operating Characteristic (ROC) curve at which the false acceptance rate and the false rejection rate are equal. The lower is the EER the better is the method. The results are reported in Table III, where it can be seen that the result obtained by the proposed ResNet-34 with contrastive-center loss is 1.7% better than the plain VoxCeleb-1024D features [9], and just 0.7% worse than the embedded VoxCeleb-256D features [9] with the advantage of not requiring any additional retraining.

D. Cross-language Speaker Verification

All the experiment conducted so far only considered the English language. In this section we want to investigate the



(a) (b) (c)

Fig. 2. ROC curves for the multilingual speaker verification test on the SIWIS dataset when the validating and claimed identity samples are: random across all languages (a), from the same language (b), and from different languages (c)

TABLE III
SPEAKER VERIFICATION RESULTS ON THE VOXCELEB AND SIWIS DATASETS IN TERMS OF EER

EER	**VoxCeleb** (%)	**SIWIS** (%)
GMM-UBM [2]	15.0	–
I-vectors + PLDA [5]	8.8	–
VoxCeleb-256D embedding [9]	**7.8**	–
VoxCeleb-1024D (reimplemented)	10.3	5.6
VoxCeleb-1024D + contr.-center loss	9.8	5.5
ResNet-18-1024D + contr.-center loss	8.6	5.4
ResNet-34-1024D + contr.-center loss	8.5	**5.2**

robustness of the proposed solution to multilingual speaker verification. The last column of Table III reports the results on SIWIS database, taking the validating and claimed identity samples randomly across all the four languages available in the dataset. A more in depth analysis of the results of the best solution proposed, i.e. ResNet-34 with contrastive-center loss, shows that it performs better when the validating and claimed identity samples are from the same language (EER equal to 3.4%) than when they came from different languages (EER equal to 6.0%). This difference in performance is not surprising, but is small enough to permit to use a model trained on a single language in a multilingual scenario. The corresponding ROC curves are reported in Figure 2.

As a further analysis we want to investigate how the model trained on English performs when tested on the other languages available (i.e. French, German, and Italian). The corresponding ROCs are reported in Figure 3.

From the ROCs it is possible to see how the model trained on English can be applied with very good results to other languages, not requiring any further adaptation.

IV. CONCLUSION

In this paper we addressed the problem of speaker recognition in the wild. We proposed the modification of two ResNet architectures, originally thought to be applied to fixed size images, to spectrograms namely ResNet-18 and ResNet-34. The proposed architectures have been trained with a contrastive-center loss in order to simultaneously obtain good results on the two tasks of speaker identification and verification, without requiring any further training or adaptation. The experimental results on the VoxCeleb and SIWIS datasets have shown the effectiveness and robustness of the proposed solution even when tested in a multilingual scenario. As a future work we plan to perform an end-to-end training without passing from handcrafted features (i.e. spectrogram based on STFT), or by injecting them in the deeper layers of the network [18].

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. The research leading to these results has received funding from The Home of Internet of Things (Home IoT), CUP: E47H16001380009 — Call "Linea R&S per Aggregazioni" cofunded by POR FESR 2014–2020. Portions of the research in this paper used the SIWIS Database made available by the University of Geneva, Switzerland.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.

[4] P. 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.

[5] 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.

[6] 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.

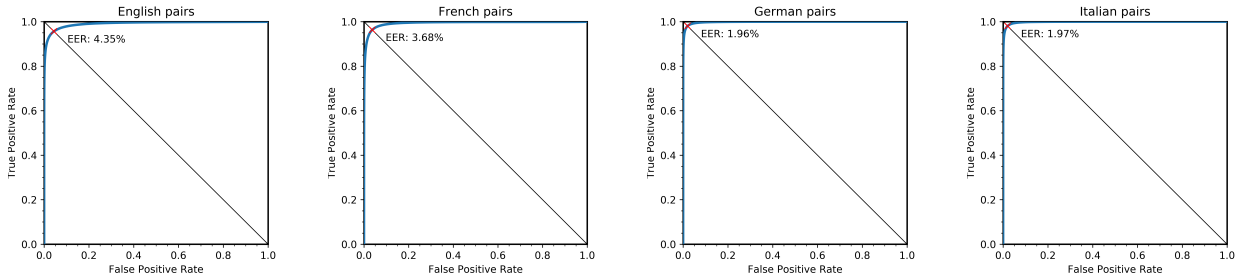


Fig. 3. ROC curves for the multilingual speaker verification test on the SIWIS dataset when the validating and claimed identity samples are: English (a), French (b), German (c), and Italian (d)

[7] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[8] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*. IEEE, 2016, pp. 1–6.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[10] 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.

[11] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[12] 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.

[13] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[14] 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.

[15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[16] C. Qi and F. Su, "Contrastive-center loss for deep neural networks," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 2851–2855.

[17] J.-P. Goldman, P.-E. Honnet, R. Clark, P. N. Garner, M. Ivanova, A. Lazaridis, H. Liang, T. Macedo, B. Pfister, M. S. Ribeiro *et al.*, "The SIWIS database: a multilingual speech database with acted emphasis," in *Proceedings of Interspeech*, no. EPFL-CONF-218855, 2016.

[18] S. Bianco, "Large age-gap face verification by feature injection in deep networks," *Pattern Recognition Letters*, vol. 90, pp. 36–42, 2017.