Text-independent speaker identification applying a very deep CNN and transfer learning

Eduardo Santos-Mena[1], Ismael de la Rosa-Vargas [2] and Aldonso Becerra-Sánchez [3]

Abstract

Speaker identification (SID) is an important discipline in the field of computational linguistics and also has several application areas. In speaker identification and speaker verification tasks traditionally, Gaussian mixture models (GMM) and vector quantization (VQ) has been applied. More recently, i-vector based frameworks are widely used by the state-of-the-art, and also interesting results have been obtained in methodologies that combine i-vectors and deep learning. This paper describes a pre-trained very deep convolutional neural network (CNN) (inception-v3) to the task of text-independent speaker identification. In this proposal the CNN works as a feature extractor and as a classifier at the same time using simple spectrograms as image-fixed inputs making our method novel, simpler and less expensive computationally compared to other recent works. The proposed metodology is evaluated on three datasets: VoxCeleb, ELSDSR, and a two local corpus with multiple recording conditions and proves to be competent against traditional GMM and i-vector methods.

I. INTRODUCTION

In computational linguistics, natural language processing has been of interest in the industrial and academic sectors with applications in areas such as bank security, personal services, criminalistics, home automation, web applications and language learning. Advances have been made applying different machine learning methodologies, in particular, numerous efforts have focused on deep learning due to substantial error rates reduction in voice recognition and related fields in recent years. In computational linguistics, there are the sub-fields: speech recognition, language recognition, speaker recognition and so on. The speaker recognition systems are divided into speaker verification (SV) and speaker identification (SID), both under a *dependent* or *independent-text* scheme. The main goal of speaker identification is to automatically infer the identity of a speaker from an input utterance given a closed set of known voice models [1, 2, 3]. Usually, in a traditional speaker identification system, a classifier is trained using acoustic features (i.e. mel-frequency cepstral coefficients) of several labeled or unlabeled audio samples in order to capture specific speaker characteristics. Different machine learning recognition methodologies have been applied with good performance in the speaker identification task which include Gaussian mixture models (GMM) [4], hidden Markov models (HMM) [5], vector quantization (VQ) [6], support vector machines (SVM) [7], Gaussian mixture model-universal background models (GMM-UBM), i-vector [8], and deep neural networks [9, 10, 11]. Specifically, the classic GMM-based method was inspired by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities [9].

The rest of the document is organized as follows. Section 2 discusses the fundamental math to the speech signal as a linear time-invariant systems and related work, in Section 3 we describe the features extraction processes, followed by Section 4, where the proposed methodology is presented, Section 5 presents the results and comparison followed by conclusion in Section 7.

II. RELATED WORK

According to the literature reviewed, several frameworks have been developed for deep learning applications in tasks of speaker identification.

Takashi Nose and Takao Kobayashi [**taka**] developed a speaker-independent HMM-based system that incorporates the use of adaptive quantization of the fundamental frequency. This was done by quantizing the average log F0 (pitch) value of each phone using the global mean and variance calculated from the training data. The study showed that with only ten sentences of the target speaker's adaptation data outperforms the conventional GMM-based using parallel data of 200 sentences. This research suggests that the use of HMM under the applied context has a better performance than conventional GMM, also a smaller amount of data is used in the training stage.

Honglak Lee et al [**Honglak**] present the use of convolutional networks of deep belief in the task of feature extraction. Specifically they use convolutional deep belief networks to audio data and empirically evaluate them on various audio classification tasks. In the case of speech data, the investigation show that the learned features correspond to phones / phonemes. Finally they evaluated the features obtained using standard supervised classifiers, such as SVM, GDA, and KNN. This work demonstrates the usefulness of convolutional neural networks in the extraction of features in speech processing, although the application of them in classification tasks was not performed.

[1]Student of the program Doctorado en Ciencias de la Ingeniería with the Department of Ingeniería Eléctrica, Universidad Autonoma de Zacatecas, ZAC 98000, MEX SantosM@IEEE.org

[2]Ismael de la Rosa-Vargas IEEE-senior member is with the Department of Ingeniería Eléctrica, and is the head of the program Doctorado en Ciencias de la Ingeniería, Universidad Autonoma de Zacatecas, ZAC 98000, MEX Ismaelrv@IEEE.org

In other hand, Egor Malykh et al [**Egor**] argue that until 2017 deep learning approaches were not very common in the speaker verification field. They investigate the possibility of using deep residual convolutional neural network with spectrograms as an input features in the text-dependent speaker verification task. As they report i-vector systems are well-known for being state-of-the-art solutions to the text-independent speaker verification task, despite the fact that they were not able to surpass the baseline system in quality, they achieved results that report an5.23% ERR for this new approach. Although this is a good approximation, it should be noted that the CNN architecture was not as deep as the one proposed in this paper.

An investigation very related to this paper is performed by Yanick Lukic et al. Considering that deep learning, especially in the form of convolutional neural networks (CNNs), has triggered substantial improvements in computer vision and related fields in recent years [**YANICK**]. They expose the application of CNNs in tasks of speaker clustering arguing that it is still common to find in the literature developments with GMM-MFCC methodologies. So in their research they face the development of an optimal design for a CNN problem, fed by simply spectrograms. The result they exhibit is compatible with those shown in the state of the art. It is important to highlight that the presented convolutional neural network has a depth of 8 layers with an architecture: convolution-pooling-convolution-pooling-dense-dropout-dense-softmax.

Another interesting work in the use of CNNs is presented in the study that is conducted by [**Christian˙bartz**], where CNNs are fed with audio spectrograms containing pronunciations in different languages. The detection of languages plays an important role in the recognition of speech in general since it helps to identify what kind of contextual and grammar rules should be applied. This research applies a hybrid convolutional recurrent neural network (CRNN) that operates on spectrogram images of the provided audio snippets. They demonstrates, that the model is applicable to a range of noisy scenarios and can easily be extended to previously unknown languages, while maintaining its classification accuracy.

Zhengwei Huang et al in 2014 apply semi-CNN for the recognition of emotions. This is done based on a mixed methodology between a system of extracting features in a unsupervised stage (CNN) fed by spectrpgrams of different resolutions and kernels with multiple sizes and a classification stage performed by SVM fed by the vector obtained in stage one. An interesting point raised in this research is how semi-CNNs are used in order to learn affect-salient features that outperforms several well-established speech emotion recognition features and their method performs an stable and robust recognition [**ZHENGWEI**].

An important research about large-scale speech recognition is presented by [**Tara N. Sainath**] et al, they apply the CCNN as an alternative to the traditional neural networks, since the speech signal has spectral variations and model spectral correlations and since CNNs have the ability to deal with these phenomena, they hypothesize that CNNs are a more effective model for speech compared to Deep Neural Networks (DNNs). This research performs experiments with the number of convolutional layers in order to determine an optimal architecture, likewise determine which is an appropriate number of hidden units and which is the best pooling strategy. Feature-space maximum likelihood linear regression features (fMLLR) are used as the input of the CNNs in order to reduce variability of speech due to different speakers. The results reported represent 12%–14% relative improvement in WER over a strong DNN achieving state-of-the art results.

Finally, Dimitri Palaz et al present the use of CNN in a speech recognition system. The author describes in general the work of an automatic speech recognition system in two fundamental stages: the feature extraction and and classifier training, which corresponds to the acoustic modeling of the phonemes as a result. In this work scheme it is shown as in the first two convolutional layers the CNN learns (in parts) and models the phone-specific spectral envelope information of 2-4 ms speech. Given that they shows that the CNN-based approach yields ASR trends similar to standard short-term spectral based ASR system under mismatched (noisy) conditions, with the CNN-based approach being more robust [**Dimitri Palaz**].

III. PROPOSED METHODS

A. Data corpus

The proposed methodology was tested in three databases: VoxCeleb1 [12], ELSDSR[13] and two local databases that together contain a wide variability between their specifications. The model created was tested in more than 10 hours of recording with an approximate gender distribution of 55% men and 45% women with sampling frequencies of 8kHz, 14.7kHz and 16kHz. Audio samples contain a variation of different sentences in the Spanish and English languages, and there are accents mainly from the countries: Mexico, Denmark and the United States. The recording condition is studio (clean) and in the wild with stereo and mono channels although in this studio only one channel (monophonic) is used. For more information refer to the table III-A

B. Inception-v3 model development

The inception v3 model was mounted on a 64-bit linux operating system in the Python programming environment on an AMD Ryzen 7 2700 processor at 3.20GHz, 8-Core. An NVIDIA GeForce RTX 2060 SUPER graphics card with 6GB GDDR6 and 2176 NVIDIA CUDA cores was incorporated on which processes are executed in parallel in order to reduce the computational load, finally a 16GB RAAM memory is used.

	Corpus			
Specifications	Voxceleb	ELSDSR	Corpus-local	Corpus-local2
f_s	16kHz	16kHz	14.7kHz	8kHz
Number of speakers	7000	22	10	12
National distribution	Multi-national	91% Danish	100%Mexican	100% Mexican
Gender distribution	61%m-39%f	46%m-54%f	60%m-40%f	50%m-50%f
Hours	+1,000	0.5	1.2	0.83
Utterances	+100,000	-	1,000	14,000
Language	Multi-languaje	English	Spanish	Spanish

TABLE I

DATASET TECHNICAL DESCRIPTION

C. Pre-procesing

The pre-processing was applied to each audio sample in the training and test data and it was performed in five main stages: normalization, voice activity detector, spectrum equalization, the spectrogram estimation, and a image-fixed transformation.

First, the waveform amplitude of the audio files was normalized to values from -1 to 1. Then, a voice activity detector (VAD) was applied to the signal in rectangular windows of 50ms, the windows with less than 20% of the maximum value of the logarithm of the energy were rejected. Pre-emphasis filtering was applied to the signal according to identity (in its discrete form) $1-\alpha Z^{-1}$ where the value of alpha is $0.94 \leq \alpha \leq 0.97$, in this study we pick $\alpha = 0.94$. Then, the spectrogram is estimated by applying the fast Fourier transform algorithm with a length in the Hamming window function of 45ms and an overlap of 10ms (for more details see [**spectrogrampaper**]). The spectrogram $S(t, f)$ (in the time frequency domain) required an additional adjustment made when operating the Equation 1

$$S'(t, f) = \log_{10}\left(S(t, f) + 1 * 1000\right), \tag{1}$$

the gray level values 0-255 were assigned as the Equation 2 indicates

$$S(t, f)_{unit8} = S'(t, f) - argmax S'(t, f) \frac{255}{argmax S'(t, f) - argmin S'(t, f)}, \tag{2}$$

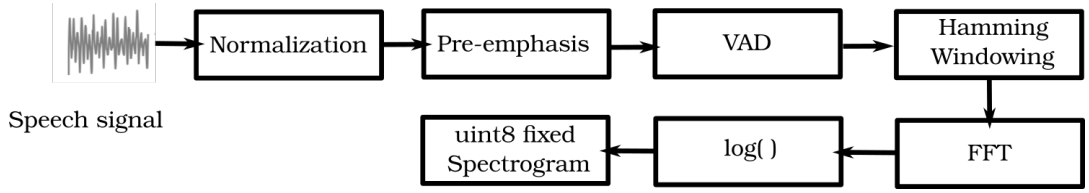

Fig. 1. Pre-procesassing metodology

D. Convolutional Neural Networks

Traditional CNNs for speech recognition consist of several convolutional layers and pooling layers, followed by several fully-connected layers for acoustic modeling. The additional convolutional and pooling layers are the main differences of CNNs compared to DNNs. A convolutional layer does the convolutions on feature maps of the previous layer using filters, and then adds a bias scalar to the corresponding feature map, followed by a non-linear operation. Feature maps are the basic units of convolutional layers and pooling layers. The typical speech inputs, with static, delta and double delta features, can be represented as 3 feature maps and each of them can be viewed as an image-map with a size of windows analsysy t_w length by number of filters f_m in the bank, usually 11x40, where t_w is the context window size of the input features and f_m is the dimension of the frequency-based features. FBANK features have been shown to be more effective than MFCC and PLP for CNN usage in speech processing, due to two reasons: 1) there is correlation information at the frequency scale represented in the FBANK features which can be utilized by convolution operations, and 2) each pooling operates on ordered-frequency feature maps, which decreases the resolution in a meaningful way. In contrast, the DCT transform of the MFCC will break this property. In figure 2 is shown the CNN applied to spectrogram.

E. Model development

The pre-trained Google CNN architecture Inception v3 is implemented in this work. This CNN used 1.28 million images and 1000 classes for its pre-training, achieving an accuracy of 93.33% on the 2014 ImageNet Challenge [INCEpTIOn]. For

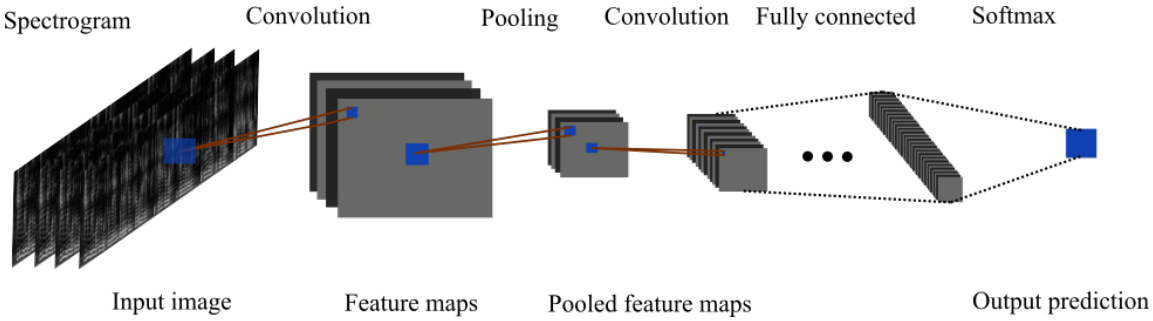

Fig. 2. CNN structure applied

this approach, the CNN is trained using a "transfer learning" technique [cite28], where the final classification layer from the network presented in figure was retrained for 5000 epochs with the elsdsr dataset. The rest of the layers are fine-tuned using the original learning parameters [cite27]. Inception CNN was chosen due the fact that such CNN can be exported for low cost hardware such as Raspberry Pi and can be deployed in Android platforms; furthermore, the CNN architecture acted as multiple convolution filters that were then applied to the same input; the results were then concatenated and passed forward. This approach allowed the model to take advantage of multi-level feature extraction. The CNN Inception V3 is based on a pattern recognition network, and it is designed to use minimal amounts of image pre-processing. Each of the proposed CNN layer reinforces key features; the first layer detects edges, and the second tackles the overall design, among others [cite29]. The Google TensorFlow deep learning framework is used to retrain the CNN; the experimental parameters were set to 5000 epochs.

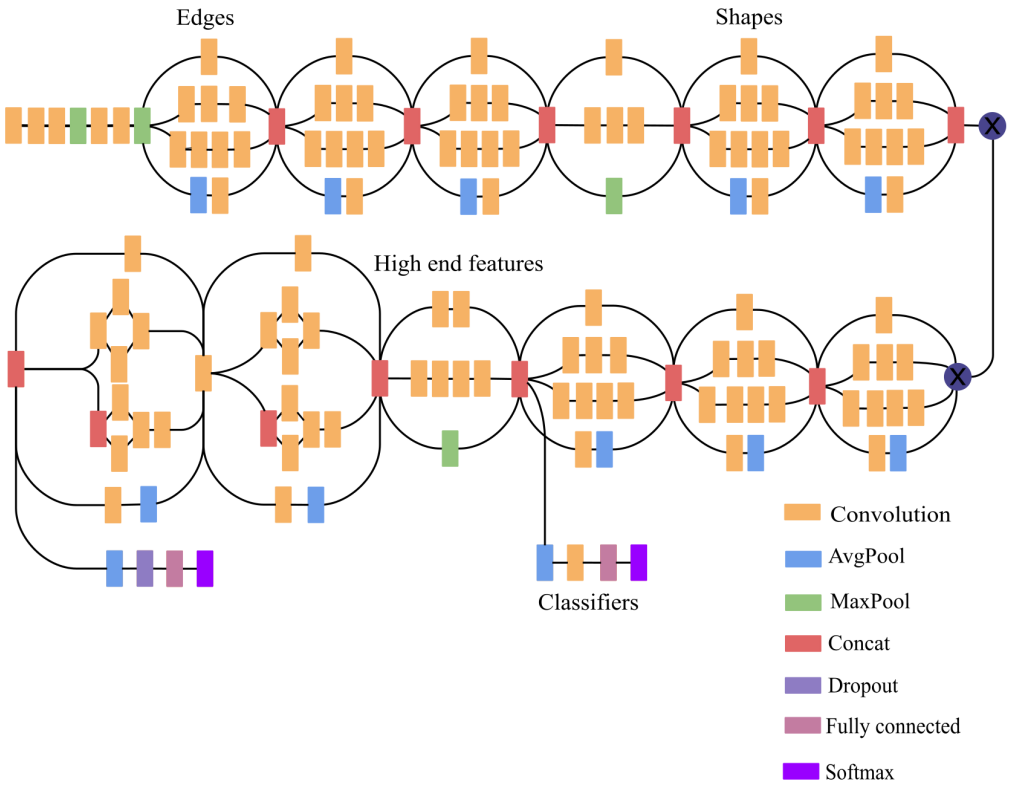

Fig. 3. Adapted Inception V3 CNN architecture

F. Definition of GMM

Let us take the multi-diementional form for the definition of a probability density function

$$N(x|\mu,\Sigma) = \frac{1}{2\pi^{d/2}\sqrt{|\Sigma|}}\exp(\frac{-1}{2}(x-\mu)^T\Sigma^-1(x-\mu)) \tag{3}$$

where μ is the mean and Σ is the covariance matrix of the Gaussian. So we define the GMM as the probability distribution that contains two probability distributions of the features vector. Then, given the feature vectors as a dataset $X = \{x_1, x_2, x_3, ..., x_N\}$, where the distribution is unknown, it must be estimated the parameters θ, maximizing the likelihood $p(X|\theta)$ as

$$\theta = argmax_\theta p(X|\theta) = argmax_\theta \prod_{i=1}^{N} p(x_i|\theta). \tag{4}$$

For the two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ there are 5 paramaters to adjust, 4 of the them are normal distributions means and standar deviation and 1 more for the probability of choosing one of them or mixture weight. Let w be the probability that the data comes from the first normal distribution, the parameter in this model is $\theta = (w, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ and the probability density function (PDF) is

$f(x|\theta) = w\ f_1(x|\mu_1, \sigma_1^2) + (1-w) f_2(x|\mu_2, \sigma_2^2)$.

In order to adjust the θ parameters, it will be randomly initiated the cluster centers and a iterative procedure refine the parameters based on an expectation-maximization algorithm defined as i) expectation step: assign each data point X_i to cluster C_i with the following probability

$$P(X_i \in C_k) = p(C_k|X_i) = \frac{p(C_k)p(X_i|C_k)}{p(X_i)} \tag{5}$$

ii) maximization step: estimation of new parameters

$$m_k = \frac{1}{N} \sum_{i=1}^{N} \frac{X_i P(X_i \in C_k)}{\sum_j P(X_i) \in C_j}, \tag{6}$$

then the iterative process continue finding the min likelihood for X and finding better parameters estimator.

IV. RESULTS

The combination of the values in the variation of results due to the change between the parameters of extraction of characteristics and in the definition of the statistical models accommodates a wide variety of results in question to the percentage accuracy of recognition by word and speaker. The corpuses used are described in table **??** and in table **??** the parameters for MFCC feature extraction are described (note: for MDFL feature extraction same configuration are used but CC are invariably).

V. CONCLUSIONS

ACKNOWLEDGMENT

REFERENCES

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

[2] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE circuits and systems magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[3] M. R. Hasan, M. Jamil, and M. G.R.M. S. Rahman, "Speaker identification using mel frequency cepstral coefficients," *Variations*, vol. 1, p. 4, 2004. [Online]. Available: `https://www.researchgate.net/profile/Golam_Rabbani4/publication/255574793_SPEAKER_IDENTIFICATION_USING_MEL_FREQUENCY_CEPSTRAL_COEFFICIENTS/links/55f05d5908ae0af8ee1d1894.pdf` (visited on 10/08/2016).

[4] 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.

[5] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell System Technical Journal*, vol. 62, no. 4, pp. 1075–1105, 1983.

[6] F. K. Soong, A. E. Rosenberg, B.-H. Juang, and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," *AT&T technical journal*, vol. 66, no. 2, pp. 14–26, 1987.

[7] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.

[8] 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, 2010.

[9] N. N. An, N. Q. Thanh, and Y. Liu, "Deep CNNs with Self-Attention for Speaker Identification," *IEEE Access*, 2019.

[10] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[11] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[12] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *ArXiv preprint arXiv:1706.08612*, 2017.

[13] L. Feng, "Speaker Recognition, Informatics and Mathematical Modelling," *Technical University of Denmark, DTU*, 2004.