DeepVoice

Rémi Hutin, Rémy Sun, Raphaël Truffet
{Remi.Hutin, Remy.Sun, Raphael.Truffet}@ens-rennes.fr
Département informatique, ENS Rennes
Campus de Ker lann, Bruz, France

Guillaume Gravier, Vedran Vukotić
{guig, vedran.vukotic}@irisa.fr
Linkmedia project, IRISA
Campus de Beaulieu, Rennes, France

***Abstract*—To carry out a speaker recognition task, i.e., to identify the speaker in an audio signal, one must first obtain a serviceable representation of said signal. Improvements upon the raw numeric signal have been made over the years, one of which is the creation of a Gaussian mixture model supervector representing the probabilistic distribution of the signal's spectral features. Due to the enormous size of such supervectors, condensed forms such as i-vectors have been created to provide usable data for application tasks. Deep learning techniques have seen significant success in many classification tasks due to their ability to build upon successive layers of data representation. We will strive to explore the possibility of using such techniques to acquire a representation of supervectors that is at least competitive with i-vectors.**

***Index Terms*—Speaker recognition; Deep learning; i-vector; Latent representation**

I. Introduction

Speaker recognition refers to the all too important identification of a speaker from an unlabeled audio signal. However, the resolution of this essential issue is no trivial matter as it raises a number of questions along the way. The most immediate one, is the manner in which the data will be represented: raw numeric signal, Fourier transform, spectral representation?

Years of research into the issue have yielded a solution called *supervectors* which provide a probabilistic representation of the numeric signal's spectral features. Those supervectors however present the significant downside of being enormous and containing information not relevant to the task at hand, which makes them unfit for applications. What recent works have had success with is the extraction of more condensed representations from those supervectors named *i-vectors*.

Deep learning techniques have had tremendous success in a number of fields ranging from computer vision [8] to natural language processing [2] by building upon successive layers of data representation and inferring hierarchical dependencies within the data. It seems to us that such architectures would be especially adapted to the extraction of intermediate representations from *supervectors* and might provide representations that outperform the current state of the art *i-vectors*.

To begin with, we will give a brief exposition of signal representations used up until now II. Afterwards, a brief explanation of the Deep Learning techniques we will need will be provided III. Finally, we will give an overview of the study we plan to conduct IV

II. Sound signals

An audio signal used for speaker recognition tasks is technically an analogical signal. To allow computer processing, those signals are sampled into a numeric signal (discrete set of measures that can a discrete set of values). Such a representation is highly vulnerable to noise and transformations however, and more developed means of representation are usually used.

A. Cepstral analysis

The speech signal of a speaker is hardly suitable for statistical modeling or the calculation of a distance. In order to obtain a representation which is more compact and less redundant, we use a cepstral representation of the speech. The cepstrum of a signal $x(t)$ is defined by (1)

$$C(x(t)) = \mathcal{F}^{-1}(\ln(|\mathcal{F}(x(t))|) \tag{1}$$

where $\mathcal{F}$ is the Fourier transform. We can analyze a signal locally by applying a short-term window. Then, we extract a vector of cepstral coefficients of this part of signal. We repeat this operation for several windows, until the end of the signal is reached.

In the cepstral domain, the distance between two speech signals may be easily computed using the Euclidean distance.

B. Gaussian mixture model (GMM) and supervectors

A Gaussian mixture model (GMM) is a probabilistic model used to approximate a distribution of random variables as a sum of normal distributions ([1]). Here, we suppose that cepstral vectors of a signal follow a probability distribution that is specific to the given signal. This distribution is the one that we try to approximate with the GMM using a reference model. In fact, even if the distribution is specific to the given signal, the form of the distribution is universal for the natural language and called the Universal Background Model (UBM). The supervector is the vector that gather means of all the normal distributions of the GMM. Such models are used to represent signals in speaker recognition tasks (Figure 1)

C. I-vectors

A supervector presents a probabilistic representation of a signal's spectral features. However, these supervectors are enormous, which makes them unfit for applications and, since they contain information on the entire signal, contain information of little interest to us. That's why we want a more condensed version a these supervectors.

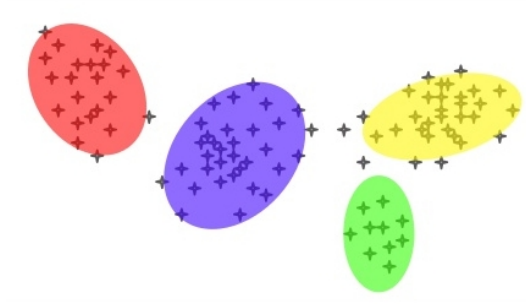

Figure 1. Gaussian Mixture Model with 4 Gaussian distributions (GMM)

Let M be the supervector, obtained as explained previously. M depends on the speaker and on the channel. We will perform a *linear projection*, to express M as

$$M = m + Tw \tag{2}$$

where

- m is a speaker and channel independent supervector
- T is a rectangular matrix, called *Total-variability matrix*
- w is an intermediate vector, or *i-vector*

The typical size of w is 60. These linearly extracted *i-vectors* provide lighter, and therefore more usable, data for application tasks.

D. Previous works

Extracting the i-vector is interesting for speaker recognition because it is easier to apply usual classifications method on it as cosine distance scoring (CDS) and support vector machines (SVM) [11, 3]. I-vectors are good tools for speaker recognition because they are channel independent and compact. We will try to keep these two characteristics in the new representation we will build using neural networks.

III. Use of neural networks

The recent success of deep learning techniques in various fields such as computer vision [8] and natural language processing [2] has sparked many explorations of their usefulness in classification tasks on *i-vectors* such as speaker recognition. Many architectures such as deep belief networks [5],[4], deep neural networks [5],[4], recurrent networks [10] or even a mix of deep neural networks and support vector machines [9] have been tested and yielded better results than previous techniques, though – to the best of our knowledge – none directly tackled the issue of supervectors' intermediate representation. Instead, all of the aforementioned work relied on pre-existing i-vector extraction techniques. We will explore the possibility of extracting a meaningful intermediate representation of supervectors through the use of deep architectures.

A. Formal neuron

a) Neuron: Neural networks can be fairly complicated to understand all at once, and therefore we will start by explaining how one of its basic units work. A neuron (Figure 2)

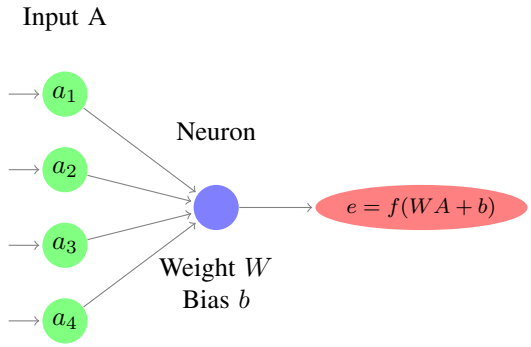

Figure 2. Formal neuron

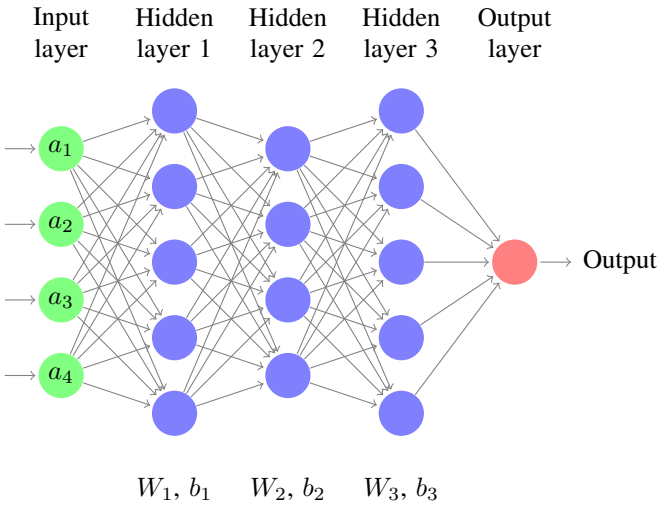

Figure 3. Deep neural network

can be thought of as a function which takes an n-dimensional vector A as input and returns a scalar e as output. This function typically has two internal parameters which are a bias b and a weight matrix W. The function starts by calculating $WA + b$ before using a non-linear activation function (such as sigmoid or tanh), i.e.,

$$e = f(WA + b). \tag{3}$$

b) Adjusting the function: Our end goal is to have the neuron, and by extension the neural network (which we will introduce later), perform a certain task. The formal neuron "learns" by adjusting its function to perform better on a designated task. For simplicity's sake, we will first explain how a process – called "back-propagation" – works with a single neuron.

For instance, suppose we have a bi-dimensional vector given as input and that we want our neuron to return 1 if its two coordinates are the identical and -1 if they are not. A natural way to evaluate how accurate our neuron consists in looking at the distance between its output e and the desired result r :

$$d(e) = |e - r| \tag{4}$$

We want to modify our neuron/function to minimize this distance. That means changing e, typically by gradient descent on function d derivative. Here, $\frac{\partial d}{\partial e} = r - e$, which means we want to "move" e in this direction. To this end, we modify our function's two internal parameters W and b. e, and by extension d, can actually be seen as a function of those two parameters: $d(W, b) = |e(W, b) - r|$. Therefore $\frac{\partial d}{\partial W} = \frac{\partial d}{\partial e}\frac{\partial e}{\partial W}$, $\frac{\partial d}{\partial b} = \frac{\partial d}{\partial e}\frac{\partial e}{\partial b}$. We then only need to compute new internal parameters W' and b' with 5 and 6.

$$W' = W + s\frac{\partial d}{\partial W} = W - s\frac{\partial d}{\partial e}\frac{\partial e}{\partial W} \tag{5}$$

$$b' = b + s\frac{\partial d}{\partial b} = b - s\frac{\partial d}{\partial e}\frac{\partial e}{\partial b} \tag{6}$$

What we just demonstrated was a simple back-propagation algorithm called gradient descent. This method is deeply flawed, but most state of the art back-propagation methods find their origins in this humble algorithm.

c) Neural network: Typically, a neural **network** (Figure 3) is made up of more than a single neuron. A neural layer refers to multiple neurons working on the same input (or parts of the same input) and producing an output that can be construed as some form of concatenation of their respective outputs. This output can be in turn regarded as an alternate representation of the input. Back-propagation for each neuron works the same way as it would if it were the only neuron calculating.

The notion of **deep** learning comes from the fact that the alternate representation computed by one layer A can be fed as input to another layer B. This allows networks to infer multiple levels of representation, same as one first processes simple geometrical forms before recognizing more complex compositions. Back-propagation is straight-forwardly computed on layer B. It is computed on layer A by looking at $\frac{\partial d}{\partial input_B}$ instead of $\frac{\partial d}{\partial e}$ which is simply calculated using the rule of chain derivation:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial x_1}...\frac{\partial x_k}{\partial x}. \tag{7}$$

What is of particular interest to us is the intermediate representations the network acquires this way. Indeed, our goal is to obtain a representation of supervectors that outperforms the linear extraction of i-vectors.

B. Autoencoders

An autoencoder [6] is a neural network with a particular architecture which is defined below.

a) A default task for a different goal: Neural networks' ability to extract internal meaningful representation from the original raw data is very enticing. Indeed, providing a meaningful representation of the initial data is one of the biggest hurdles of classical machine learning. However, neural networks need to be trained on a task, and require labeled data to that end.

Since we are no longer interested in the specific task needed to train the network, it is possible to default to a task that is not interesting in itself but might require the network to learn salient features of the data in order to be completed. A very natural one is to match the output to the input, which would not require man made labeling. Hopefully, the learned representation will capture features specific to the input so that it may be reconstructed.

b) Structure: Typically, the **input** is *encoded* (8) into a **latent representation** which is then *decoded* (9) into an **output** that should match the input as closely as possible, hence the autoencoder denomination (Figure 4). The main danger of such an approach is having the networks that does not transform the input at all and yields a latent representation that is no different to the raw input data. A multitude of solutions exist to avoid such a scenario:

- Compress the input into a latent representation of lower dimension
- Put the input through a corrupting input layer: the autoencoder cannot just do nothing and give back the same thing because it never had the actual input to begin with. This particular structure is called denoising auto-encoder.
- Use various regularization methods.

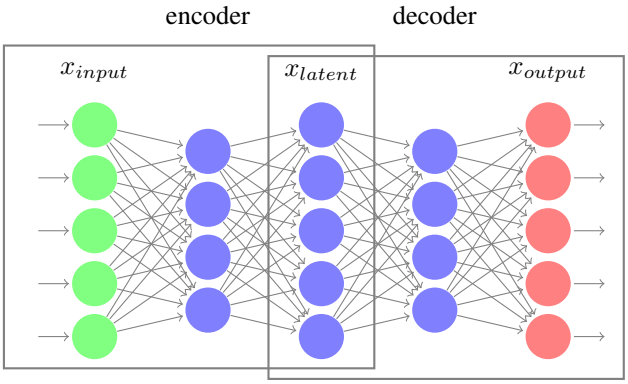

Figure 4. Structure of an autoencoder

$$x_{latent} = encoder(x_{input}) \tag{8}$$

$$x_{output} = decoder(x_{latent}) \simeq x_{input} \tag{9}$$

c) Why use autoencoders: Autoencoders can be used for denoising, using corrupted data as input and the original data as objective. Autoencoders may also be very useful to learn a representation. In fact, the latent representation contains enough information to reconstruct the data.

In our context, autoencoders could be used to find a representation of a speaker, by giving to the autoencoder a supervector from a speaker as input, and an another supervector from the same speaker as objective. This idea is based on considering the within-speaker variability as a noise, and using a denoising autoencoder. Repeating this with several pairs of supervectors, we may learn a representation of the speaker. This new representation, as well as i-vectors, is more condensed than supervectors.

Since the problem is symmetrical, we can perform the learning in both directions. One way to perform this training is to use an autoencoder that reconstruct each supervector from the other using tied weights.

C. Tied weight autoencoder

We talk about tied weight autoencoder when the same weights appears in two different places in the autoencoder. For example, in the architecture proposed in [12], the weights in light blue in the upper part of the autoencoder are the same as the weights in light blue in the lower part. So, if the function that transform the upper first hidden layer to the upper part of the representation is

$$h_1(x) = f(W \times x + b_1) \tag{10}$$

then the function that transform the lower part of the representation to the next layer is

$$h_2(x) = f(W^T \times x + b_2) \tag{11}$$

In this architecture (see Figure 5), the pink weights are also tied.

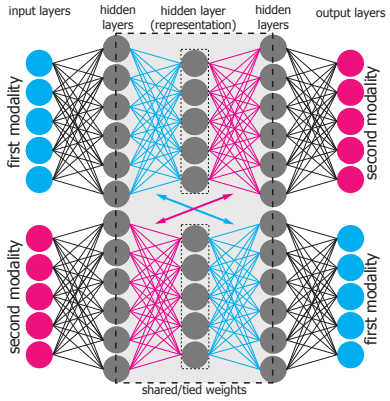

Figure 5. Symmetrical and bidirectional learning architecture using tied weights

This architecture learns a joint representation of the two modalities which gives encouraging results multi-modal query expansion [12]. In our context, the two modalities will be two supervectors from the same speaker.

IV. Method

As explained earlier, supervectors computed from Gaussian mixture models provide useful features of a given signal. i-vectors have built upon this representation to extract a more compact representation that better represent the specificities of the signal. One task that naturally comes to mind is deciding whether two signals have been said by the same person. Our suspicion is that linear extraction of i-vectors provides suboptimal results for such computations. Therefore, we will use neural networks to extract an intermediate representation of each signal that might be better suited to this particular problem.

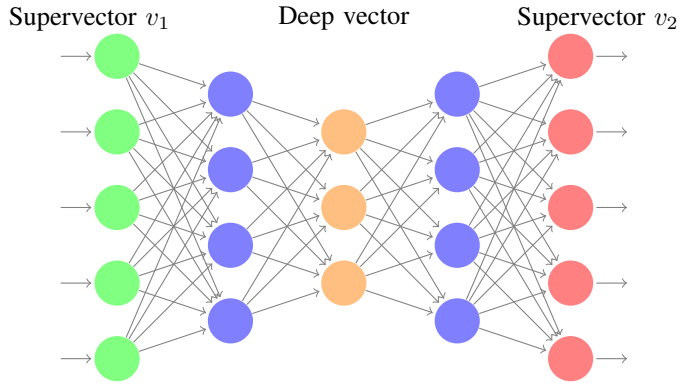

Figure 6. General prospective model

A. Prospective method

a) Phasing out non-speaker dependant noise: In order to extract a representation of supervectors that makes speaker dependent features of the given signal more salient, we use a model derived from the design of a denoising autoencoder. Indeed, we have no need for any information pertaining to what is actually being said or what kind of microphone was used: this is all noise to us. Therefore, if two supervectors v_1 and v_2 originate from the same speaker, it seems reasonable to think of them as the same original "speaker" sound vector which has polluted by non-speaker dependant noise.

The latent representation thusly extracted should present salient features specific to the studied speaker. An autoencoder trained that way could allow for a sort of projection that we wish to study.

b) Evaluating speaker-dependant similarities: We wish to acquire intermediate representations that draws different signals spoken by the same speaker closer together according to some measure of distance, therefore allowing for an easy classification. The distances we will mostly consider during our study will be angular distance and Euclidean distance.

To this end, we will have to carefully chose the distance function used to evaluate the "quality" of our output in order to make sure we will be optimizing this particular aspect of the acquisition.

c) Prospective general model: We will therefore train an autoencoder on supervectors. Given good results of [12], we will tie all but the extreme weights to reflect the symmetrical nature of the task at hand. A rough outline of the general model is shown in Figure 6.

B. First experimentation

In light of the difficulties inherent to the training of large neural networks, a preliminary study on a small dataset will first be conducted.

1) Dataset:

a) Processing raw data: 15311 audio signals were extracted from BFMTV's various programs and labeled with the name of their respective speaker. Those sound files were then processed into 15311 supervectors of length 9216 with the

AudioSeg tool. We then create 1 839 235 pairs of supervectors that originate from the same speaker.

b) Input: For each computed pair, we feed both of its supervectors to the neural network as input. Therefore, we train the network on 3 678 470 supervectors of size 9216.

c) Output and task: The neural network outputs a vector of size 9216 that we try to match as closely as possible to the supervector the original input was paired with.

2) Neural approach: We will start by training a 7-hidden layer auto-encoder. The encoder is 3 layer deep and encodes the input of size 9216 into a latent representation of size 40. This is done by successively compressing into representations of size 5000 (first layer), 500 (second layer) and 100 (third layer). The decoder is 3 layer deep and inflates the 40 dimensional latent representation back into a 9216 dimensional supervector by successively inflating into representations of size 100 (fifth layer), 500 (sixth layer) and 5000 (seventh layer). The weights used to inflate from the latent layer to the fifth and the fifth to sixth are tied to those used to compress from the second layer to the third and from the first layer to the second.

3) Evaluating results:

a) Evaluation method: We will first try to pinpoint a threshold on the distance between two vectors which separates pairs of vectors labeled to the same speaker and labeled to two separate speakers. To this end, we will compute the distance (discussed in IV-A) between every pair of vectors labeled to the same speaker. The maximum of those distances is then computed and chosen as our **threshold**.

We know by construction that every pair labeled to the same speaker is below the chosen threshold. To evaluate the relevance of the studied representation we then count the number of pairs of unmatched vectors whose corresponding distance is below our threshold. and would be misclassified if our threshold is was used as a classification method.

b) Comparing to i-vectors: To substantiate our claim that i-vector extraction does not lead to a clustering of same speaker vectors, we will apply the evaluation method outlined above to both our extracted intermediate vectors and standard *i-vectors*. Those intermediate representations will be extracted from source files using the tool ALIZE [7].

C. Anticipated issues

a) Optimization function: It is likely we will end up tweaking the network's optimization function over the course of our study, be it by using a linear combination of the functions we already brought up or by using a completely different function.

b) Training set: We intentionally started out with a fairly small training set in order to be able to put forth some preliminary results faster so that early decisions can be made. It is possible we will not have enough training examples to train very deep architectures and will end up using bigger training set.

c) Evaluation method: The evaluation method detailed above is fairly clumsy in that it is extremely vulnerable to outliers. Various improvements on it could be explored if we notice that we consistently have some outliers that artificially skew the results in favour of one type of intermediate representation.

d) Network hyper-parameters: The above hyper-parameters (number of layers, dimension of each layer, ...) have mostly been chosen arbitrarily. The values chosen seemed reasonable in light of the pursued goal and the task at hand, but will most likely evolve depending on experiments' results.

APPENDIX

REFERENCES

[1] Frédéric Bimbot et al. "A Tutorial on Text-Independent Speaker Verification". In: *EURASIP Journal on Advances in Signal Processing* 2004.4 (2004), p. 101962. ISSN: 1687-6180. DOI: 10.1155/S1110865704310024. URL: http://dx.doi.org/10.1155/S1110865704310024.

[2] Antoine Bordes et al. "Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing." In: *AISTATS*. Vol. 351. 2012, pp. 423–424.

[3] Najim Dehak et al. "Front-end factor analysis for speaker verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798.

[4] Omid Ghahabi and Javier Hernando. "Deep belief networks for i-vector based speaker recognition". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, pp. 1700–1704.

[5] Omid Ghahabi and Javier Hernando. "Deep Learning for Single and Multi-Session i-Vector Speaker Recognition". In: *CoRR* abs/1512.02560 (2015). URL: http://arxiv.org/abs/1512.02560.

[6] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507. ISSN: 0036-8075. DOI: 10.1126/science.1127647. eprint: http://science.sciencemag.org/content/313/5786/504.full.pdf. URL: http://science.sciencemag.org/content/313/5786/504.

[7] Anthony Larcher et al. "ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition." In: *Interspeech*. 2013, pp. 2768–2772.

[8] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[9] Fred Richardson, Douglas Reynolds, and Najim Dehak. "Deep neural network approaches to speaker and language recognition". In: *IEEE Signal Processing Letters* 22.10 (2015), pp. 1671–1675.

[10] George Saon et al. “The IBM 2016 English Conversational Telephone Speech Recognition System”. In: *CoRR* abs/1604.08242 (2016). URL: http://arxiv.org/abs/1604.08242.

[11] Mohammed Senoussaoui et al. *An i-vector Extractor Suitable for Speaker Recognition with both Microphone and Telephone Speech*.

[12] Vedran Vukotic, Christian Raymond, and Guillaume Gravier. “Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications”. In: *ICMR*. ACM. New York, United States, June 2016. URL: https://hal.inria.fr/hal-01314302.