ref?

PROJ: DeepVoice

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

Guillaume Gravier
guig@irisa.fr
Linkmedia project, ~~INRIA~~ Irisa
Campus de Beaulieu, Rennes, France

+ Vedran Vukotić

Abstract—**To carry out a speaker recognition task, ie to identify the ~~locutor~~ 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 ~~modeling~~ 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.**

i.e.,

?G

I. Introduction

Speaker recognition refers to the all too important identification of a locutor 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, fourrier transform, spectral representation?

F

Years of reasearch 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 therefore evolving in a very sparse representation space, 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 ([7]) to natural language processing ([1]) by building upon successive layers of data representation and inferring hierachical 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

verb?

II. Sound signals

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

→ means of what?

A. Cepstral analysis

The speech signal of a locutor 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 :

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

where $\mathcal{F}$ is the Fourier transform. We can analyze a signal locally by applying a short-term window, ~~whose duration is shorter than the signal~~. Then, we exctract a vector of cepstral coefficents ~~of this part of signal. We repeat~~ it 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. Here, we suppose that cepstral vectors of a signal follow a probality 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.

=> refs! + usage in speaker recognition

C. I-vectors

Ajouter les références

A supervector presents a probabilistic representation of a signal's spectral features. However, these supervectors are enormous, which makes them unfit for applications. That's why we want a more condensed version a these supervectors.

not only

Let M be the supervector, obtained as explained previously. M depends on the speaker and on the channel. We want to express M as

i-vector = linear projection (factor analysis)
→ notion of separating speaker characteristics from other factors.