1 Introduction

There are more and more situations arising where people need a system to store or to exchange their personal informations. The most used solution in such a case, is to use secret codes or personal cards. This is the case for bank accounts, computer passwords, etc. The drawback of these traditional systems occurs when the secret code is lost or stolen. Who never wrote his personal identification number (PIN) code or password somewhere in order not to forget it? An alternative solution is to use biometric information, such as fingerprint, face, iris or voice, that are expected to be somehow unique to each individual, in order to restrict the access of a service to registered clients only. This is called "biometric authentication". In this thesis, we address the problem of biometric authentication using some pre-recorded human voices using an automatic system based on machine learning algorithms.

1.1 What is a Speaker Verification System?

A speaker verification system should verify the claimed identity of a person based on his voice. Basically, it has to accept as a *client* or reject as an *impostor* a speaker that claimed an identity. Different systems can be considered:

- Text-dependent systems: the phonetic content of the pronounced sentence is fixed. For example, the system can ask the speaker to pronounce a specific sentence.
- Text-independent systems: the phonetic content is free.

The former has the advantage to be robust to "replay" attacks (when an impostor plays back a pre-recorded sentence pronounced by the real speaker), but has the drawback that it needs more complex models and is very strict

about the sentence pronounced by the speaker. In this thesis, we will consider only text-independent speaker verification systems that are the most used for their simplicity, as they do not require complex speech recognition modules and they are thus better adapted to various embedded applications (phone, personal digital assistant, etc.)

While speaker verification systems have been researched and developed in the last 20 years, it is only more recently that they have benefited from research in machine learning thanks to the computational power of modern computers. Before describing the objectives of this thesis, let us first explain what is machine learning.

1.2 What is Machine Learning?

Machine learning is a research domain at the crossroad of computer science and statistics that consists in developing algorithms that allow computers to improve, "learn", automatically through experience. In order to "learn" a solution to a problem, the algorithm needs some "training" examples corresponding to this problem for which the solution is known. The overall goal is then to find the best function over a selected set of functions, according to a given loss function applied to the training examples. The set of functions should be rich enough to contain a good solution but simple enough in order to "generalize" the concepts underlying the training examples to new, unseen, examples. The size of the chosen set of functions is directly related to a formal concept know as the *capacity* of a set of functions; the solution found by a machine learning algorithm is called a *model* and the set of training examples is called a *dataset*. The machine learning community developed several algorithms that can be applied to various problems, such as speaker verification, face detection, text categorization, etc.

1.3 Road Map of the Thesis

The aim of this thesis is to address the speaker verification problem from a machine learning point of view.

A common problem in machine learning is to classify examples into two categories, the so-called two-class classification problem (Bishop, 1995). The common approaches used to solve a two-class classification problem are either discriminant (trying to find an hyperplane that best separates the two given classes) or not, the latter being often implemented using generative models

REF C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

(that try to estimate separately the distribution of each class, then relying on Bayes rule to take a decision). According to Vapnik (2000), one should never try to solve a more complex task than the one at hand. Hence, discriminant models should be favored in general. In this thesis, we consider speaker verification as a two-class classification problem for each client, with one class representing the client and the other representing the impostors.

First, we present the text-independent speaker verification systems in Chapter 2 as found in the literature and we note that unfortunately most results are usually presented using biased measures. Chapter 3 thus describes new unbiased measures and statistical tests in order to compare objectively the different proposed approaches. We used machine learning principles to design new speaker verification databases and protocols in order to produce unbiased results. In Chapter 4, we describe all benchmark databases used in this thesis to compare our new approaches to the state-of-the-art systems.

Looking at the current speaker verification literature, it is interesting to note that the dominant state-of-the-art model does not appear to be discriminant as it is based on generative models. In fact, the devil is in the details, as the speaker verification community proposed many modifications in order to reach state-of-the-art performance. When we analyze the resulting system more deeply, as done in Chapter 5, we can see that, due to these modifications, the state-of-the-art system becomes discriminant. But in this case, why not use directly discriminant models? In Chapter 5, we also propose a new generic framework that also includes discriminant models for speaker verification. We also extend this framework to score normalization techniques, that are used to make the decisions taken by a system more robust with respect to different recording conditions.

The speaker verification problem has some specificities that make the application of discriminant models difficult. First, the examples are encoded as variable length sequences of multi-dimensional vectors that depend on the phonetic content of the pronounced sentence and the speech rate of the user. Unfortunately, most discriminant models can only work on fixed size vectors. In Chapter 6, we address this problem by using informations taken from estimated densities in order to produce fixed size vectors that can be used by discriminant models. In Chapter 7, we then propose to use instead a particular discriminant model, called Support Vector Machine (SVM), which projects the examples into a high dimensional space before trying to discriminate the two classes. This projection is done using a specific mathematical function

REF V. N. Vapnik. *The nature of statistical learning theory*. Springer, second edition, 2000.

called "kernel" that can in theory be tailored to any kind of data structure. We thus propose in Chapter 7 a new kernel that can handle sequences. The recent speaker verification literature also proposed "sequence" kernels, called Generalized Linear Discriminant Sequence (GLDS) kernels that are limited to a polynomial form. Our approach gives a new enlightenment to these kernels and extend them to other kernel functions such as Gaussian kernels. As our approach is costly for long sequences and thus not applicable in some cases, we propose an approximate method to reduce its complexity.

An other particularity of the problem is that the number of positive examples (coming from the client) and the number of negative examples (coming from the impostors) are highly unbalanced. Indeed, each time the system enrolls a client, it needs records coming from this client. It is usually not realistic, from the application point of view, to ask a client to pronounce sentences several times per day during several months. We thus have only few (often only one) accesses to enroll a client. As we do not have "real" impostor accesses, the records coming from other speakers are used as negative examples, which can be several hundreds. In summary, we have a number of two-class classification problems equal to the number of clients to enroll, with about one positive example and hundred of negative examples for each problem. Fortunately we observed empirically that for all SVM based approaches the problem is separable and thus, as the SVM considers only examples in the margin, the ratio between negative and positive examples is reduced and the solution found by the SVM seems good. Instead, we address an other problem which we consider more important: the variance of the intra-client distance distribution is more peaky than the variance of the intra-impostors distance distribution. We thus propose, in Chapter 8, to create a new similarity measure by modifying the kernel by adding a Gaussian noise distribution around each negative example. Unfortunately, in order to obtain good performance, we have to modify a nice principled approach. Even if the final approach has not yet a good theoretical justification, it is a good starting point for future research.