Article

A Text-Independent Speaker Authentication System for Mobile Devices

Florentin Thullier [1,*], Bruno Bouchard [1] and Bob-Antoine J. Ménélas [1,*]

[1] Department of Computer Science and Mathematics, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada

* Correspondence: florentin.thullier1@uqac.ca; bob-antoine-jerry_menelas@uqac.ca

Academic Editor: name
Version August 2, 2019 submitted to Cryptography

Abstract: This paper presents a text-independent speaker authentication system for mobile devices. A special attention was placed on delivering a fully operational application, which admits a sufficient reliability level, as well as an efficient functioning. To this end, we have excluded the need for any network communication. Hence, we opted for the completion of both the training and the identification processes directly onto the mobile device through the extraction of Linear Prediction Cepstral Coefficients (LPCCs) and the Naïve Bayes algorithm as classifier. Furthermore, the authentication decision is enhanced to overcome misidentification through access privileges that the user should attribute to each application beforehand. To evaluate the proposed authentication system, eleven participants were involved in the experiment, conducted in both quiet and noisy environments. Moreover, we also employed public speech corpora to compare this implementation to existing methods. Obtained results have shown that our system is reliable and efficient enough in real use cases. Moreover, we suggest that the proposed authentication system should also be either employed as part of a multilayer authentication, or as a fallback mechanism.

Keywords: Speaker Authentication; Text-Independent; Mobile Devices; LPCCs; Naïve Bayes; Voice; Security

1. Introduction

Nowadays, mobile devices take a significant place in people's everyday life. As announced by the Gartner Institute, smartphone sales surpassed one billion units in 2014 [1] and people have their mobile devices everywhere and at any time [2] since they are considered as an important part of their life [3]. As a result, users do store private data such as pictures, videos, as well as secret information (*i.e.* emails, bank account) on their devices. However, people are, most of the time, not adequately wary about the safety of these secret information [4]. Within a mobile device context, authentication remains the first entry point for security. Indeed, such a mechanism aims at protecting the digital identity of users.

Over the past few years, various authentication schemes have been proposed. Mobile devices often only offer authentications that involve recalling a piece of information such as PIN code. However, they concede several drawbacks. As an example, it was reported that half of the population do not lock their phone at all [5]. They estimate that entering a PIN code involves lots of inconveniences every time the mobile device has to be unlocked [5]. Moreover, it is known that users have trouble remembering all passwords they use nowadays [6]. It is clear that these behaviors lead a huge impact on the security of mobile devices. Accordingly, people's authentication usages may generate serious threats to the security that a system initially provides [7–9]. Recently, biometric authentication mechanisms such as fingerprint, ear shape or gait recognition were enabled on mobile devices [10–13]. These systems chiefly rely on the uniqueness of the user's physiological or behavioral trait. In the same way, speaker authentication refers to the process of accepting or rejecting a speaker that claims identity. Such schemes are also advised as biometrics since they focus on vocal characteristics produced by the speech and not on the speech only. These features depend on the dimension of the vocal tract, mouth, and nasal cavities, but also rely on voice pitch, speaking style, and language [14].

Speaker authentication systems may be designed according to two leading methods: text-dependent and text-independent [15,16]. A text-dependent authentication involves the user to pronounce a predefined pass-phrase that is considered as a voice password. It is used both for the enrollment and the identification process. For instance, *Google* recently introduces the trusted voice feature on *Android*, where users have to enroll their voice by pronouncing "*Ok Google*" three times. Then, this pass-phrase must be repeated each time the mobile device needs to be accessed.

By contrast, text-independent schemes are able to identify the user accurately, for any delivered word or locution. Presently, existing speaker authentication techniques offered on mobile devices always require network communications. Indeed, matching templates are usually stored in the cloud. In that sense, they may represent costly authentication solutions for certain users. However, since no additional sensors are needed, they remain inexpensive solutions to authenticate users as regards hardware requirements. Conversely, since manufacturers have pushed fingerprint systems in the forefront of the mobile devices authentication mechanism scene, they tend to become usual. Nevertheless, fingerprints admit a major drawback since they are impossible to use in countries having hard weather conditions as people wear gloves in winter. In that sense, a speaker authentication approach may be a convenient way to resolve such an issue. Moreover, these authentication systems offer a sufficient acceptance rate with end-users and remains less intrusive than fingerprint or retina scan [8,17]. Additionally, these mechanisms may play a major role in some real-world applications to secure identity management systems such as e-commerce solutions, attendance systems, mobile banking or forensics.

It is known that several proposed speaker recognition and identification systems achieve accurate results [18–20]. Despite the effectiveness of these mechanisms, few of them are presently implemented on mobile devices. Indeed, they are mostly machine-centered implementations. Moreover, the considerable number of users who still do not secure the access to their mobile devices [5] reveal a need for novel methods mainly focused on a human-centered design that must take into account the diversity of user profiles and usages [21]. This research targets these needs. In that sense, the contribution of this paper is to expose the design of a a text independent speaker authentication system applied to mobile devices that tend to focus on users'needs. The choise of a text-independent solution is motivated by a relevant usage when there are social interactions. Indeed, saying "*Ok Google*" in the middle of a conversation may be disruptive while a text-independent solution is capable to identify and authenticate the owner of the mobile device all along the conversation without any care for what is being said. Moreover, a recent study [22] highlighted that 62% of the panel of *Android* users rarely employ the voice assistant feature and most of them have declared that *"they feel uncomfortable talking to their technology, especially in public"*.

The system we propose in this work is a mobile application designed to be extremely convenient for the user. It allows them to forget that they are using a voice-based authentication mechanism. In order to achieve such an authentication, our approach relies on Linear Prediction Cepstral Coefficients (LPCCs) and the Naïve Bayes algorithm for patterns classification. Since authentication schemes usually either grant, or deny the access to the whole content of the phone, we further suggest enhancing such a final decision to overcome false positive and negative identification that may occur and reduce annoying situations for the user. Therefore, we introduce the notion of access privileges that enable restricting certain access, based on a simple evaluation of the user's location and the presence of a headset. Moreover, we pay attention to produce an efficient system since we opted for low complexity algorithms and we avoid network communications by achieving both the training and the identification, on the mobile device itself.

The rest of the paper is structured as follows: Section 2 provides an overview of related work to proposed speaker identification and verification systems. Section 3 details the suggested system, specifically designed to entirely operate on mobile devices. Next, section 4 describes experiments we conducted in order to evaluate the reliability as well as the efficiency of such an implementation. Section 5 exposes and discusses results we obtained. Finally, section 6 draws a conclusion and section 7 provides future works.

2. Related Work

In order to achieve speaker authentication, several techniques have been described for years through disparate features extraction techniques and classification algorithms. This section first exposes suggested text-independent speaker identification and authentication systems to determine their suitability as regards a usage on mobile devices. Finally, we will examine proposed schemes which were explicitly designed to operate on mobile devices.

First of all, Reynolds and Rose [20] have proposed a text-independent speaker identification which exploits Mel-Frequency Cepstral Coefficients (MFCCs) as features and a Gaussian Mixture Model (GMM) to predict which person is speaking. MFCCs are widely used in speaker recognition as they accurately represent the envelope of the short-time power spectrum of the signal. Although such coefficients appear to be more robust against noisy conditions, their acquisition remains very expensive regarding mobile device capabilities [23]. The main motivation of using a GMM was based on an empirical observation that a large number of unlabeled classes of sample distribution, may be represented as a linear combination of Gaussian basis functions. To evaluate the system, a subset of the KING speech database was used. This database provides utterances from speaker conversations over both signal-to-noise radio channels and narrow-band telephone channels. An accuracy of 80.8% was obtained for 49 telephone speech samples of 15 seconds. Besides, authors claimed that this model is computationally inexpensive and easy to implement on real-time platforms. However, the main drawback of such a system lies in the initialization of the training procedure, where parameters such as mean, covariance and prior of each distribution have to fit the data. Indeed, such a process may be achieved through several costly methods like a Hidden Markov Model (HMM), or a binary k-means clustering algorithm. In that sense, although the identification process may certainly be efficient when used in a mobile device context, the training phase should probably be computationally overly expensive.

Secondly, Kumar *et al.* [18] have suggested another text-independent speaker identification approach which aims at predicting utterances thanks to a back-propagation neural network, where LPCs (Linear Prediction Coefficients) parameters were used as input features. The goal of the back-propagation method is to optimize weights between each neuron layers so that, the neural network can learn how to correctly map arbitrary inputs to outputs. Hence, outcomes provide the resulting decision in determining at which speaker corresponds each given utterance. The evaluation of the system was performed over a collection of 25 speech samples in different languages. An overall accuracy measure of 85.74% was achieved. This led authors to state that such a technique remains appropriate and reliable. However, the theoretical complexity of a standard back-propagation neural network training phase is $O(nmh^koi)$, where n are training samples, m refers to features, k are hidden layers, each containing h neurons, o refers to output neurons and i is the number of iterations [24]. Hence, such a computation time remains overly expensive in a mobile device context.

On the other hand, concerning speaker authentication, Nair and Salam [19] have proposed a text-independent system which exploits both LPCs and LPCCs to compare their strength. The decision was made through the Dynamic Time Warping (DTW) algorithm. DTW allows calculating the distance between two given sequences which provides the optimal match. Authors have experimented their system over the TIMIT speech corpus which provides 630 real speech signals of American English speakers. An overall accuracy of 92.2% was obtained with LPCs features while it rose up to 97.3% with the derivative cepstral coefficients. Combining LPCCs with the DTW algorithm involves thus, an accurate and reliable solution to authenticate users by their voice. Since DTW requires a quadratic time and space complexity (*i.e.* $O(n^2)$) [25], it may not be the most suitable method to achieve speaker authentication, directly on the mobile device. Nevertheless, real speaker authentication scenarios usually imply few distinct samples. In that sense, DTW as decision-making still stays an acceptable choice for such an authentication mechanism on limited-performance devices.

The growing interest in deep leaning approaches observed in recent years forced to question us about its suitability as regards a speaker identification task. Lee *et al.* [26] have shown that Convolutional Deep Belief Networks (CDBN) features trained with a SVM classifier have outperformed MFCC features trained with a GMM (respectively to the method described in [27] when the number of training examples was small (*i.e.* 1 to 2 utterances per speaker). However, with a greater number of training samples (*i.e.* 5 to 8 utterances per speaker) results remained similar (*i.e.* around a 99% accuracy). Moreover, since deep learning algorithms yet remain costly in terms of computational power and processing time, the training process is always achieved on the server-side [28]. However, with the recent partnership developed between *Movidius*, the leader in low-power machine vision for connected devices and *Google*, next generation mobile devices may embed a chip dedicated to the computation of complex machine learning algorithms such as deep neural networks [29]. Then, in the present situation, it appears that such an approach may not be an adequate solution according to our need.

To the best of our knowledge, it appears that few text-independent speaker authentication solutions for mobile devices were applied in practice. For instance, Vuppala *et al.* [30] have suggested a recognition model which lies in several speech enhancements to improve the overall performance faced with different noisy conditions. In that sense, authors aimed to prove the robustness of the method when used in varying background environments. However, their evaluation was performed through noise simulations over speech samples from the TIMIT corpus.

Conversely, Brunet *et al.* [31] have introduced a practical text-independent speaker authentication system which is entirely usable on mobile devices. The approach suggests extracting MFCCs features from speech samples. Then, a reference model is built thanks to a Vector Quantization (VQ) method. The Euclidean distance between stored centroids and testing samples is calculated and compared to a given threshold in order to accept or reject the attempt. Authors have performed an experiment over their own database, where training and testing samples were collected thanks to a mobile device, as well as the Sphinx database which contains 16 American English speakers'utterances. Since the method was implemented as a stand-alone biometric system, only the Equal Error Rate (EER) was computed to evaluate the performance. Hence, they obtained better performances on their database (4.52 of EER at best), than the ones on the public database (5.35 of EER at best). However, achieved results largely rely on initial parameters required for the quantization step (*i.e.* the number of centroids) that must be optimized according to the training data. With this brief review, it appears that few text-independent authentications that focus on mobile device computation capabilities and generic usages were proposed. Hence, this paper introduces a user-centered text-independent speaker authentication system for mobile devices. A special attention was paid to its usability and the effectiveness of the training as well as the identification steps in order to compute both of them directly on the mobile device. As a matter of fact, we selected low-computational cost algorithms that do not require any parameter to optimize with others expensive techniques regarding processing time as long as they may offer an accurate identification.

3. Proposed Speaker Authentication System

In this research, we introduce a text independent speaker authentication system for mobile devices. This method works as stand-alone and does not require any costly architectures such as client/server. Hence, the entire computation is done end-to-end on the mobile device. As illustrated in Figure 1, this mechanism consists of three main processes. The first one involves extracting individual voice features from a raw audio input to build a data set. The following operation lies in training such data with a Naïve Bayes classifier. The last process is the authentication decision. It aims at enhancing the conventional speaker verification mechanism to increase the confidence rate of the authentication. To achieve this, we suggest granting a specific access privilege to the user through the evaluation of two different states. The first one concerns the current location of the user that is compared to the ones defined beforehand. Secondly, we evaluate the presence of a headset, that is, whether if it is plugged in the mobile device or not. Indeed, we assume that to use a headset with a built-in microphone reducing surrounding noises will decrease chances for a user to be unwillingly authenticated on this mobile device though replay attacks. [32]

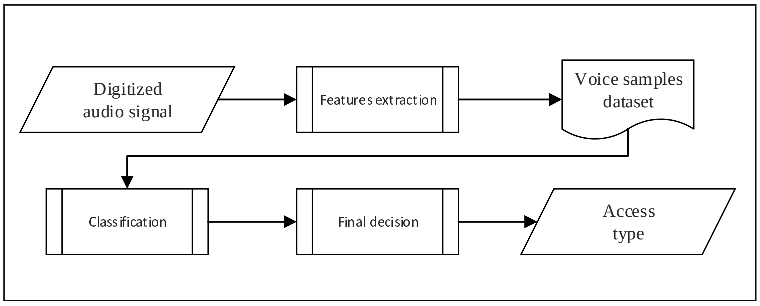


Figure 1. Flowchart of our proposed speaker authentication system.

3.1. *Input*

Audio files are recorded using a 16-bit signed integer PCM encoding format in bi-channels. The sampling rate of such audio files is up to 44.1 kHz.

3.2. *Pre-processing*

3.2.1. Voice Activity Detection (VAD)

Given an audio record as input, the first step that we produce is a pre-processing phase that aims at removing every silence area only to keep speech segments. To this end, we first have defined a given threshold close to zero (*i.e.* 0.0001). Then, our main focus in such a pre-processing step is to identify which section of the input signal is close to this threshold in order to remove it. To achieve this, we apply the autocorrelation function $r_x(t,k)$ suggested by Sadjadi and Hansen [33] onto a windowed audio segment $s_w(n)$ of the entire input signal $s(n)$ given by,

$$r_x(t,x) = \frac{\sum\limits_{n=0}^{N-1} s_w(n)w(n)s_w(n+k)w(n+k)}{\sum\limits_{n=0}^{N-1} w(n)w(n+k)}, \tag{1}$$

where t and k are frame and autocorrelation lag indices, respectively, and $w(n)$ is a Hamming window given by,

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N_w - 1}), & 0 \le n \le N_w - 1 \\ 0, & otherwise. \end{cases}, \tag{2}$$

where its length (N_w) is based on the frequency of the signal.

For each processed segment $s_w(n)$, if the mean value of the computed coefficients, resulting from the autocorrelation function, gets close to the defined threshold then, it is identified as a silence area and we finally remove it. Figure 2 graphically illustrates this process.

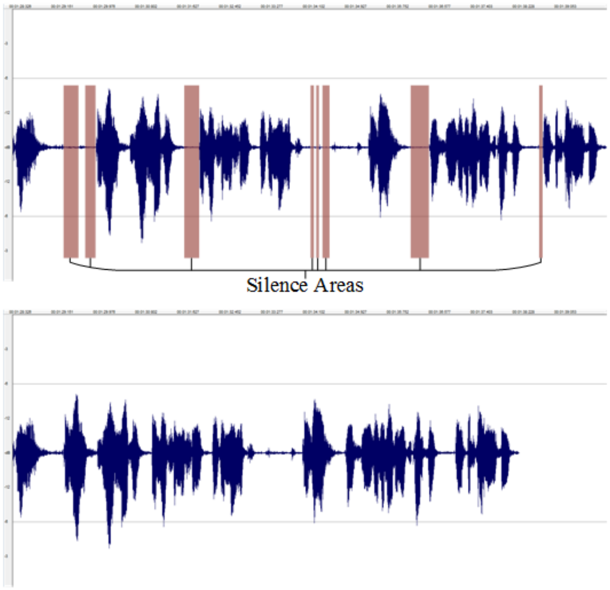


Figure 2. The first signal is the raw input where silence areas are highlighted. The second is the output of the same signal after the silence removal process.

3.2.2. Audio Normalization

Succeeding the silence removal phase, a peak normalization is performed. The goal is to change the gain of the input to the highest peak of the signal, uniformly. Traditionally, this process is used to ensure that the highest peak remains at 0 dBFS (deciBels relative to Full Scale)—the loudest level allowed in a digital system. Since the entire signal is adjusted, it is indistinguishable and does not affect the original information. Moreover, the process of normalization ensures that the audio will not clip in any manner. Figure 3 shows the graphical result of such a process.

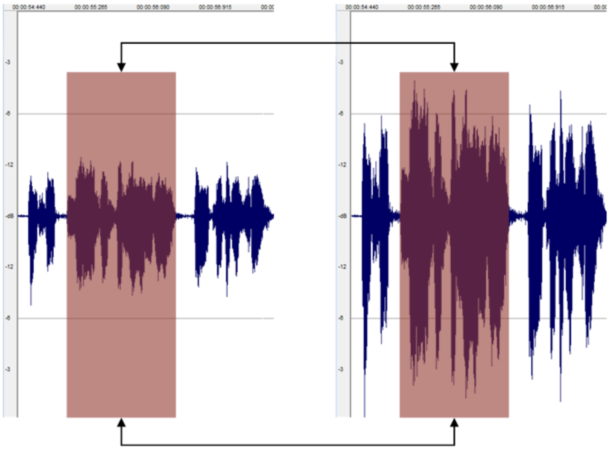


Figure 3. The left signal is the input signal and the right one is the same signal with peak normalization, where the same sequence is highlighted on both signals.

3.3. Feature Extraction

Since the voice is considered as a signal containing a lot of information about the speaker—the process of extracting several discriminating features from the speech remains a critical part of both speaker identification and authentication systems. In that sense, we decide to favor the use of the Linear Prediction Cepstral Coefficients (LPCCs). Such coefficients are directly derived from the Linear Prediction analysis that aims at estimating the relevant features or characteristics from a speech signal

[34]. We justify such a choice by its ability to provide extremely accurate estimates of the speech parameters, and by its relative speed of computation [23]. This last point was a crucial criterion since mobile devices presently remain less powerful than traditional desktop computers. Figure 4 graphically summarizes the steps of the features extraction from a pre-processed signal to the resulting data set containing voice features.

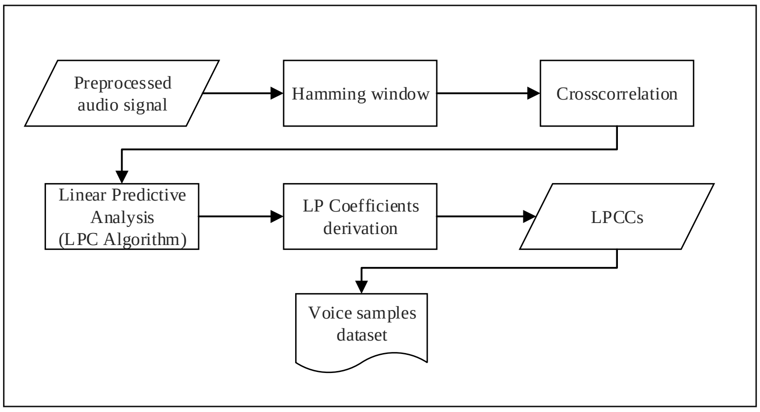

Preprocessed audio signal
Hamming window
Crosscorrelation
Linear Predictive Analysis (LPC Algorithm)
LP Coefficients derivation
LPCCs
Voice samples dataset

Figure 4. Flowchart of the features extraction process.

To compute the LP analysis, we have implemented the Linear Predictive Coding algorithm. It was designed to exploit the redundancy present in the speech signal by assuming that each sample may be approximated by a linear sum of the past speech samples (p). Hence, the predicted sample $S_p(n)$ may be represented as,

$$S_p(n) = \sum_{k=1}^{p} a_k s(n-k) \, , \tag{3}$$

where $a(k)$ are the Linear Prediction Coefficients (LPCs), $s(n-k)$ are past outputs and p is the prediction order. In our case, the speech signal is multiplied by an overlapped Hamming window of 25ms to get a windowed speech segment $S_w(n)$ as,

$$s_w(n) = w(n)s(n) \, , \tag{4}$$

where $w(n)$ is the windowing sequence given in equation (2). The error between the actual sample and the predicted one $e(n)$ may be expressed as,

$$e(n) = s_w(n) - \sum_{k=1}^{p} a_k s_w(n-k) \, . \tag{5}$$

The main objective of the LP analysis is to compute the LP Coefficients that minimize this prediction error. To this end, our system exploits the autocorrelation method that is usually preferred since it is computationally more efficient and more stable than the covariance one [35]. Thus, the total prediction error E is given as,

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} \left(s_w(n) - \sum_{k=1}^{p} a_k s_w(n-k) \right)^2 \, . \tag{6}$$

The values of $a(k)$ that minimize this total prediction error may be computed by finding,

$$\frac{\delta E}{\delta a_k} = 0, \quad 1 \leq k \leq p \, . \tag{7}$$

Thus, each a_k gives p equations with p unknown variables. The equation (8) offers the solution to find LP Coefficients,

$$\sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n) = \sum_{k=1}^{p} a(k) \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n-k), \quad 1 \leq i \leq p\,. \tag{8}$$

Consequently, it is possible to express the linear equation (8) in terms of the autocorrelation function $R(i)$ as follows,

$$R(i) = \sum_{n=i}^{N_w} s_w(n)s_w(n-i), \quad 0 \leq i \leq p\,, \tag{9}$$

where N_w is the length of the window. Then, by substituting values from equation (9) in the equation (8) with the autocorrelation function $R(i) = R(-i)$ we obtain the following equation,

$$\sum_{k=1}^{p} R(|i-k|)a_k = R(i), \quad 1 \leq i \leq p\,. \tag{10}$$

The set of linear equations is expressed by the relation $Ra = r$ and may be represented in a matrix form as,

$$\begin{array}{cccc} R & a & & r \\ \begin{bmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} & \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} & = & \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \end{array}, \tag{11}$$

where a is the vector of LP coefficients and r is the autocorrelation. The resulting matrix is a Toeplitz matrix where all elements along a given diagonal are equals.

Towards the computation of the LP Coefficient a_k, it is possible to derive cepstral coefficients c_n directly through the following relationship,

$$c_n = \sum_{k=1}^{n-1} a_k c_{n-k} + a_n, \quad 1 < n \leq p\,, \tag{12}$$

where p refers to the prediction order.

It is known that speaker recognition requires more cepstral coefficients than speech recognition which employs around 15 of them. Although it was pointed out that increasing the number of such coefficients does not affect the recognition [36], we suggest using 20 LPCCs to preserve a relatively good computation speed.

3.4. Classification

Several classification algorithms were employed for speaker recognition (*i.e.* GMM, ANN, *etc.*). However, it is known that the Naïve Bayes classifier is fast, very effective and easy to implement. As a supervised and statistical learning method for classification—it simply computes the conditional probabilities of the different classes given the value of attributes. Finally, it selects the class with the highest conditional probability. Accordingly, Table 1 exposes the theoretical time and space complexity evaluations of the Naïve Bayes classifier [37].

Table 1. Naïve Bayes time and space complexities, given k features for both training and testing operations [37].

Operation	Time	Space
Training on n samples	$O(nk)$	$O(k)$
Testing on m samples	$O(mk)$	$\Theta(1)$

Once the feature extraction process is completed, a set of samples denoted $s_1, s_2, \ldots, s_i$ with their associated class labels $c_{s_1}, c_{s_2}, \ldots, c_{s_i}$, where $c_{s_i} \in \Omega = \{c_1, c_2, \ldots, c_i\}$ is obtained. Each sample has k features (*i.e.* LPCCs) represented by floating numbers (with $k = 20$), that are denoted as $a_1, a_2, \ldots, a_n$. The objective of the Naïve Bayes classifier is to exploit these samples to build a model (*i.e.* the training phase) that will be reused to predict the label of the class c_p for any future sample (*i.e.* the identification phase). Figure 5 shows a simplified block diagram of this process.

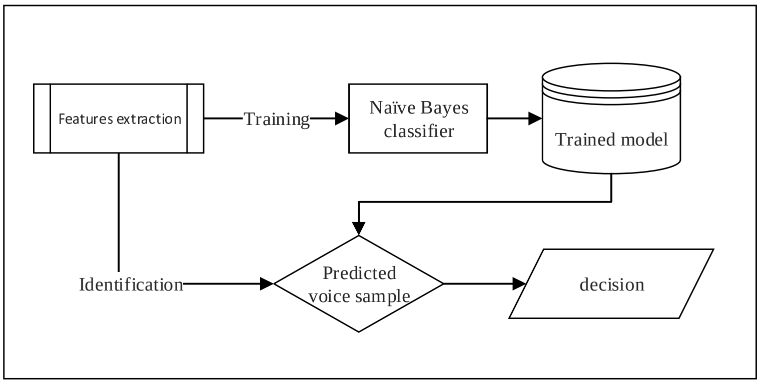


Figure 5. Flowchart of the classification process.

The algorithm, strongly relies on the Bayes theorem and imposes two assumptions. Firstly, all features $a_1, \ldots, a_n$ should be independent for a given class c. This is the class-conditional independence. Secondly, all features $a_1, \ldots, a_n$ should be directly dependent on their assigned class c. Given that, it is possible to describe the classifier as,

$$P(c|a_1, a_2, \ldots, a_n) = \frac{P(c) \prod_{i=1}^{n} P(a_i|c)}{P(a_1, a_2, \ldots, a_n)} \,. \tag{13}$$

Since $P(a_1, a_2, \ldots, a_n)$ is common for a certain sample, it may be ignored in the classification process. As a result, we can derive equation (13) to predict the class c of a given sample during the identification phase as follows,

$$c = arg \max_{c \in \Omega} P(c) \prod_{i=1}^{n} P(a_i|c) \,. \tag{14}$$

However, as we obtain the LP coefficients through an autocorrelation method, resulting LPCCs remain strongly dependent and consequently, violate the independence assumption of the Naïve Bayes classifier. Nevertheless, Zhang [38] has demonstrated that such a condition is not necessary to satisfy in practical situations. Indeed, no matter how strong dependencies among attributes are, Naïve Bayes can still be optimal if they are distributed evenly in class, or if they cancel each other out. Moreover, we have observed that the distribution of our features, for all classes, when compared to their frequency, follows a normal distribution. Hence, it is possible to assume a valuable classification rate with Naïve Bayes according to the supposed quality of the LPCCs.

3.5. Decision-making

The result provided by the classification task may return two kinds of errors. On the one hand, a false-negative outcome refers to a failure in a genuine authentication, while a false-positive result concerns an impostor attempt mistakenly identified. According to authentication on mobile devices, false-negative does not compromise the security of the private content. However, it may be disturbing for the user since either the process has to be repeated, or a fallback mechanism has to be used. In contrast, false-positive exhibits a serious vulnerability for the security of such devices since the objective is to avoid fraudulent accesses. Besides, speaker authentication systems are not devoid of other drawbacks that may also lead to security threats. Indeed, they remain vulnerable to voice mimicry or mock authentication through legitimate voice records.

Hence, we suggest improving the authentication process by introducing the notion of access privileges in order to ward against misidentification. The implementation suggested in this paper remains experimental, but it is easily possible to imagine that such a process may become a standard of mobile operating systems. Firstly, we assume that users have assigned a right for each application installed on their mobile device beforehand. Therefore, we define three privileges as follows. The public privilege allows the access to only non-critical content and applications. The protected privilege restricts the access to the most critical pieces of data (*i.e.* bank account). Finally, the private privilege gives the access to the entire content of the mobile device.

The process of determining the safest authority to grant that we suggest begins by verifying the result produced by the identification process. If the voice does not match with a genuine one then, the system allows the user to have a public access. In case of false-negative, the user has to repeat the entire process otherwise, an impostor identification is avoided. If a match does exist, a protected access is granted and the current location is fetched. In that case, the system verifies that the position is inside a given radius between 200 and 500 meters of one trusted location—where trusted locations refer to a predefined set of places connected with the user (*i.e.* home, work). This verification allows us to be quite more robust against fraudulent authentication attempts. However, a risk still exists, especially when we are facing users living together such as a family or roommates who obviously share at least, one same location. Hence, to reduce chances for a user to be unwillingly authenticated on his own device, we offer to proceed another verification. Indeed, we suggest that the private access level must only be allowed when the authentication process is achieved while using a headset and all previous verification are satisfied. Therefore, we both verify that the headset is plugged into the output of the device and that it provides an extra microphone. We justify such a need because these microphones are closer to the mouth of the speaker and they better filter the surrounding noise than the built-in microphone of the mobile device. Thus, by bypassing the mobile device microphone, we estimate that it represents an additional level of security when there are shared trusted location. In that sense, we assume that false acceptance rates must considerably decrease. Figure 6 graphically summarize this process.

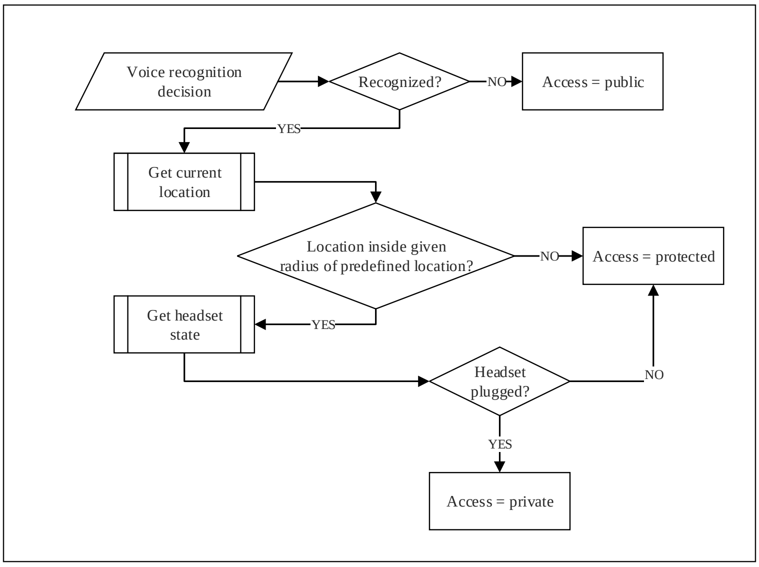


Figure 6. Flowchart of the decision-making process.

4. Experiments

The experiment we have conducted aims at assessing the proposed authentication system. Moreover, it let us suggest a public data set of 11 speakers voice features (*i.e.* UQAC-Studs). To achieve this, every participant used the system to authenticate himself or herself on the provided mobile device, with a headset plugged. Two distinct environmental conditions were exploited in this experiment. The first one was a quiet environment, where the training process and a quiet authentication were completed. Finally, another authentication attempt was performed within a noisy environment in order to evaluate the robustness of such a mechanism.

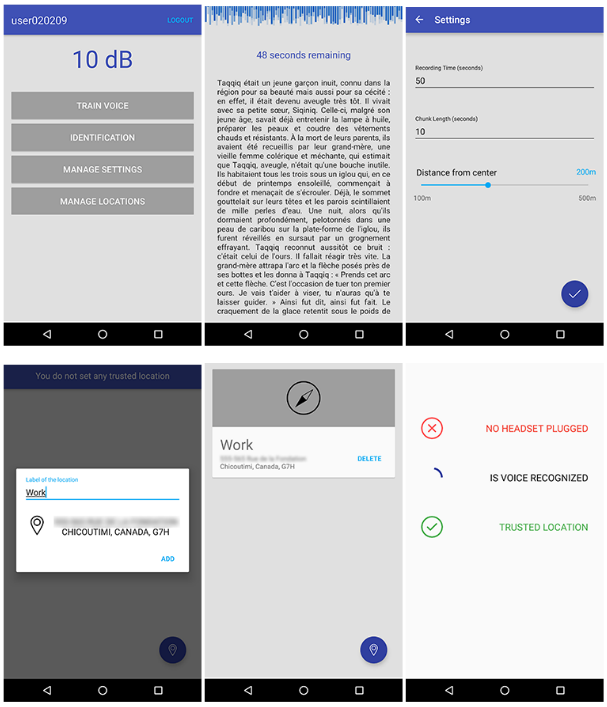
4.1. Participants

We recruited 11 university students as participants, 7 males, and 4 females from 19 to 36 years. All participants were speaking French but some distinct accent, such as Canadian French and hexagonal French were observed. Moreover, they were either *iOS*, or *Android* users and owned at least one recent mobile device (*i.e.* smartphone or tablet). Furthermore, 9 of the participants used an unlocking mechanism for their smartphone (PIN: 4, pattern: 2, fingerprint: 3) and fingerprint users either had a PIN code, or a pattern as fallback mechanism.

4.2. Data Collection

The proposed text-independent system was implemented as an *Android* application which requires at least the 4.0.1 version of the mobile operating system. Figure 7 shows screen captures of the application. All participants performed the experiment on the same smartphone (*i.e. LG Nexus 5* running *Android* 6.0.1 with a *Snapdragon 800* Quad-core at 2.3 GHz CPU and 2 GB of RAM) with the same headset (*i.e. Bose SoundTrue I*) in same conditions (*i.e.* room and public place).

Since it was desired to have real-environment recording conditions, a quiet room was selected to achieve the training, as well as the quiet identification session. Conversely, the noisy session was performed in the cafeteria of the University. The sound level of each distinct place was measured thanks to a sound level meter embedded in the application. The mean value evaluated reached 16.5 dB in the quiet environment while 95 dB was observed in the noisy one.



user020209

10 dB

48 seconds remaining

Taqqiq était un jeune garçon inuit, connu dans la région pour sa beauté mais aussi pour sa cécité : en effet, il était devenu aveugle très tôt. Il vivait avec sa petite sœur, Siqiniq. Celle-ci, malgré son jeune âge, savait déjà entretenir la lampe à huile, préparer les peaux et coudre des vêtements chauds et résistants. À la mort de leurs parents, ils avaient été recueillis par leur grand-mère, une vieille femme colérique et méchante, qui estimait que Taqqiq, aveugle, n'était qu'une bouche inutile. Ils habitaient tous les trois sous un iglou qui, en ce début de printemps ensoleillé, commençait à fondre et menaçait de s'écrouler. Déjà, le sommet gouttelait sur leurs têtes et les parois scintillaient de mille perles d'eau. Une nuit, alors qu'ils dormaient profondément, pelotonnés dans une peau de caribou sur la plate-forme de l'iglou, ils furent réveillés en sursaut par un grognement effrayant. Taqqiq reconnut aussitôt ce bruit : c'était celui de l'ours. Il fallait réagir très vite. La grand-mère attrapa l'arc et la flèche posés près de ses bottes et les donna à Taqqiq : « Prends cet arc et cette flèche. C'est l'occasion de tuer ton premier ours. Je vais t'aider à viser, tu n'auras qu'à te laisser guider. » Ainsi fut dit, ainsi fut fait. Le craquement de la glace retentit sous le poids de

Recording Time (seconds)
50

Chunk Length (seconds)
10

Distance from center 200m

You do not set any trusted location

Label of the location
Work

[redacted]
CHICOUTIMI, CANADA, G7H

Work
[redacted]
Chicoutimi, Canada, G7H

NO HEADSET PLUGGED

IS VOICE RECOGNIZED

TRUSTED LOCATION

Figure 7. Screen captures of the *Android* application.

4.3. Procedure

In the beginning, participants were introduced to the experimental procedure and the current position was added to the trusted location list.

Then, training participant voices was the first phase of the experiment. To complete such an operation, a text was randomly selected in a database and displayed on the screen of the device. Participants were instructed to wear the headset and to familiarize with the content. Once they were ready, participants were advised to start the recording by themselves and next, to begin reading the text aloud. The record was automatically stopped after one minute by the application and participants were warned through both a vibration and a text-to-speech synthesis system. At that point, participants were asked to wait until the end of the computation. In the meantime, the main recorded file was split into 10 seconds'chunks, being 6 instances per class in total. Each set of features from each instance were written in the data set which was used to create the training model of the Naïve Bayes classifier, as described previously. Finally, participants were advised of the completion of the process thanks to a pop-up message.

At the end of the training process, the authentication process starts. This procedure was performed twice. In the first place, participants were asked to wear the headset and to pronounce the locution of their choice in the quiet environment. In the second place, they were requested to execute the same task in the noisy environment. Insofar as there was no restriction on the locution which had to be said—participants were able to use either two different expressions, or the same one for the two authentication sessions. Since every authentication attempts were performed in the same place, our decision-making has always stated that users stood in a trusted location. Therefore, we have mocked a location which was not considered as a trusted one afterwards, in order to verify the reliability of our technique. Figure 8 summarizes the proceedings of the experiment we conducted using a sequence diagram.

Finally, in the last step of the experiment, participants were sounded out about their habits concerning authentication on their own device, as well as their opinion as regards the proposed system.

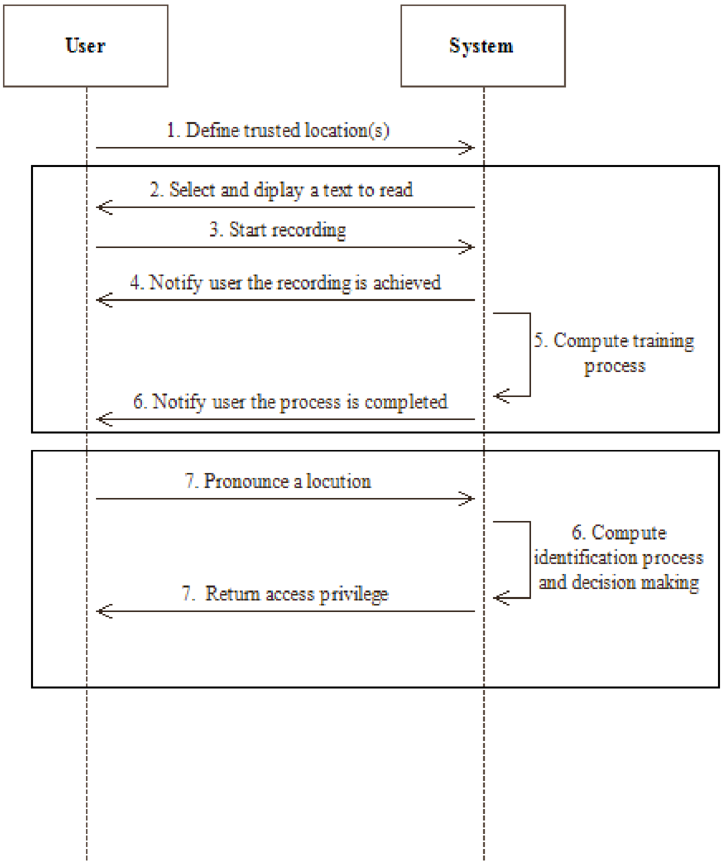

User
System
1. Define trusted location(s)
2. Select and diplay a text to read
3. Start recording
4. Notify user the recording is achieved
5. Compute training process
6. Notify user the process is completed
7. Pronounce a locution
6. Compute identification process and decision making
7. Return access privilege

Figure 8. Sequence diagram of the experiment.

5. Results and Discussion

5.1. Speech Corpora

In this research, we have evaluated the performance of our system by exploiting two additional speech corpora for comparison purpose with the data set we suggest. The first one is the Ted-LIUM corpus which has been proposed by Rousseau *et al.* [39]. It includes a total of 1495 audio files extracted from TED talks, where all speeches are English-based with multiple distinct accents. These records are mono-channel, and they are encoded in 16-bit signed integer PCM at a 16 kHz sampling rate. Although the corpus was published using the NIST Sphere format (SPH), we required to convert the whole files in Waveform Audio File Format (WAV). Furthermore, we took care of removing the first fourth frames of each file, as they correspond to the talk opening sequence. The second speech corpus that we have exploited in this research is a subset of the TIMIT corpus, which has been suggested by Garofolo *et al.* [40]. Such a subset contains 10 broadband recording files for 16 English-based speakers. Such provided files are also mono-channel and encoded in 16-bit integer PCM at a 16 kHz sampling rate.

5.2. Classification Performance Metrics

Since classification let us predict at which registered speaker corresponds a given utterance; it is important to evaluate the performance of our system thanks to representative metrics. To this end, the accuracy is probably the most dominant measure in the literature, because of its simplicity. This measure provides the ratio between the correct number of predictions and the total number of cases given as,

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{15}$$

where TP and TN refer to true positive and true negative predictions respectively, and the total additionally include false positive (FP) and false negative (FN) predictions.

Despite its popularity, accuracy alone does typically not provide enough information to evaluate the robustness of prediction outcomes. Indeed, accuracy does not compensate for results that may be expected by luck. Indeed, a high accuracy does not necessarily reflect an indicator of a high classification performance. This is the accuracy paradox. For instance, in a predictive classification setting, predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy. In that sense, as suggested by Ben-David [41], we decided to provide the Cohen's kappa evaluation metric as well. This measure takes into account such a paradox and remains a more relevant metric in multiclass classification evaluations such as our system. The kappa measure is given by,

$$kappa = \frac{P_o - P_e}{1 - P_e}, \tag{16}$$

where P_o and P_e are the observed and the expected probabilities respectively.

5.3. Results Obtained

The performance of our proposed system was evaluated according to several analyses. First of all, results of the experiment we described previously are shown in Table 2. In this evaluation, we have exploited testing instances we obtained over our experiment for both quiet and noisy environments. Thanks to such achieved results it is possible to observe that our system yields an acceptable identification of voices in real environmental conditions with our instances.

Table 2. Results of the experiment based on the realized data set: UQAC-Studs.

	Quiet environment	**Noisy environment**
Accuracy	91%	82%
Kappa	90%	80%
Total classes	11	11
Total instances for training	5	5
Total instances for identification	1	1

However, since it is impossible to state the reliability of the results we obtained with only such data, we have constructed related data sets thanks to the Ted-LIUM and the TIMIT subset corpora as a means of comparison for our system. For all 16 speakers, the TIMIT subset admits ten recorded files between two and four seconds. Hence, we have exploited 6 samples to construct the training set and the four remaining were used for the identification. Results we obtained over this speech corpus are shown in Table 3.

Table 3. Results obtained over a subset of the TIMIT speech corpus.

Accuracy	83%
Kappa	82%
Total classes	16
Total instances for training	6
Total instances for identification	4

Nevertheless, since the Ted-LIUM speech corpus is large and contains several long records, we judged that it was a necessity to unify the construction of the data sets according to the previously described subset of TIMIT corpus. In that sense, we have created ten different training sets by selecting 16 samples randomly over the 1495 files. Moreover, we have also ensured that a sample was not chosen more than once for a given batch. For each batch of ten records, every sample is split into 10 instances of 5 seconds. In order to be more consistent with our experimental procedure, the first 6 instances are used in the training phase; while the last four are exploited for the identification. Figure 9 details the results obtained for these ten random batches. In addition, such an experiment has revealed a mean accuracy of 87% and a mean kappa measure of 85%.

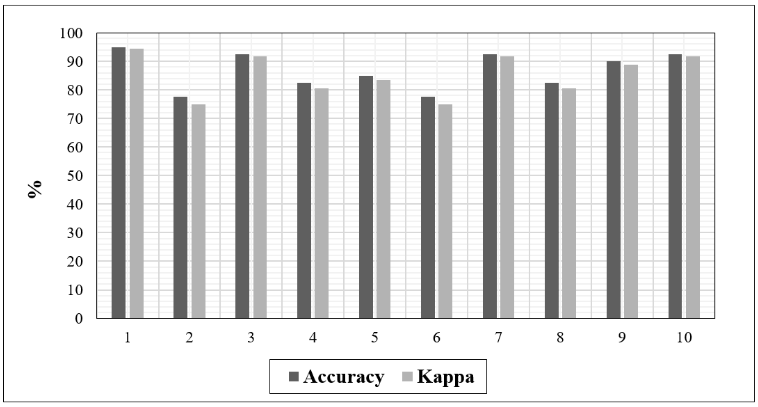


Figure 9. Accuracy and kappa measures achieved by our system over the 10 random batches of the Ted-LIUM corpus we have created.

However, as these evaluations involve a relatively small number of distinct classes, we point out the analysis of the evolution of the kappa measure when increasing the number of classes. The Ted-LIUM corpus let us perform such an appraisal since it is the largest corpus we used in this research. Hence, we did not change the number of instances that we have exploited in the previous evaluation, six instances per class for the training and four for the identification phase. We chose to compute the kappa by increasing the number of classes exponentially until reaching the closest value to the total of 1495 records. Figure 10 shows that the more there are classes, the more the kappa measure tends to decrease. Indeed, our system obtains a kappa of 47% where the entire set of classes was used in the identification process. Such a result was expected since we are not facing a binary classification problem.

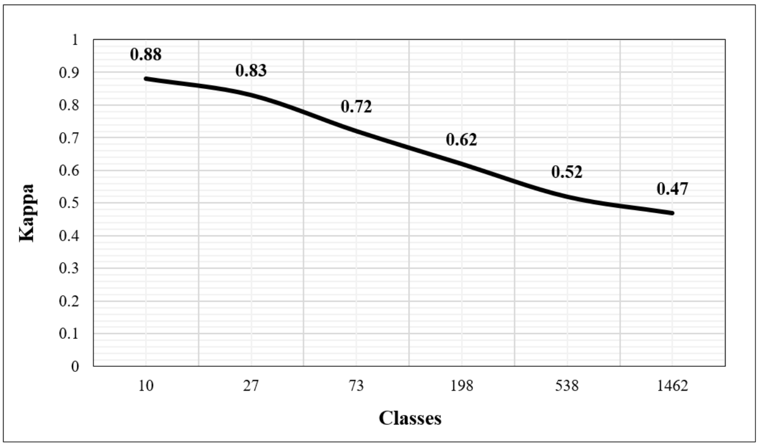


Figure 10. Evolution of the kappa measure over the Ted-LIUM corpus when increasing the number of classes exponentially.

Finally, an empirical comparison between our proposed method and previous works is exposed in Table 4.

Table 4. Empirical comparison between our text-independent speaker authentication and previous works.

	Features	**#Features**	**Classification, Pattern Matching**	**Accuracy**	**Dataset**
Thullier *et al.*	LPCCs	20	Naïve Bayes	91%∼82% 87% *(avg)* 83%	UQAC-Studs Ted-LIUM TIMIT (subset)
Nair and Salam [19]	LPCs & LPCCs	20, 30 & 40	DTW	90.4% *(20 LPCs)* 94.8% *(20 LPCCs)*	TIMIT
Reynolds and Rose [20]	MFCCs	100 12-dimensional vectors per second	GMM	96.8% 80.8%	KING Private samples
Kumar *et al.* [18]	LPCs, LPCCs, RC, LAR, ARCSIN & LSF	N.A.	ANN *(backpropagation)*	85.74%	Private samples

5.4. Replay Attacks

Replay attacks refer to the presentation of a recorded audio sample of a genuine voice played-back to get access to the protected system [32]. Since this kind of attack is considered to be the major security drawback of voice-based authentication mechanisms, it is relevant for us to state about the robustness of our system as it stands. Indeed, no specific method to counteract replay attacks such as [42] has been implemented in this work.

In order to proceed such an evaluation, the testing instance, for each participant of our experiment, was replayed to the authentication system through a standard desktop computer speaker. As expected, six utterances over the eleven were genuinely identified without the headset. However none fraudulent samples were correctly identified while using the headset which embeds its own microphone.

5.5. Computation Performances Considerations

Since we desired to create a user-centered text-independent speaker authentication mechanism, we judge that an efficient, as well as a reliable implementation is an important angle when considering to replace most used and weak authentication mechanisms such as PIN codes.

To this end, we have chosen suitable techniques with attention to time complexity and memory consumption. Figure 11 exposes a profiling of CPU, memory and battery utilization of the mobile device, in relation to the cumulative time consumption. These measurements were performed on every stage of the training, as well as the identification processes, for one given instance of the data set we suggest. Moreover, the start stage was considered as an idle state for the application.

In order to produce six instances of ten seconds each, we had to record during 60 seconds. Hence, the whole training process has required less than ten seconds of processing, while the identification process has demanded less than 500 ms to terminate since we recorded during two seconds. Moreover, the memory usage did not exceed 70 MB.

These measurements were observed through the *Trepn Profiler* application developed by *Qualcomm*, but since accurate performance metrics are difficult to obtain, it is impossible for us to provide a suitable analysis of the battery needs. Nevertheless, we only present a trend of the required power consumption for the application.

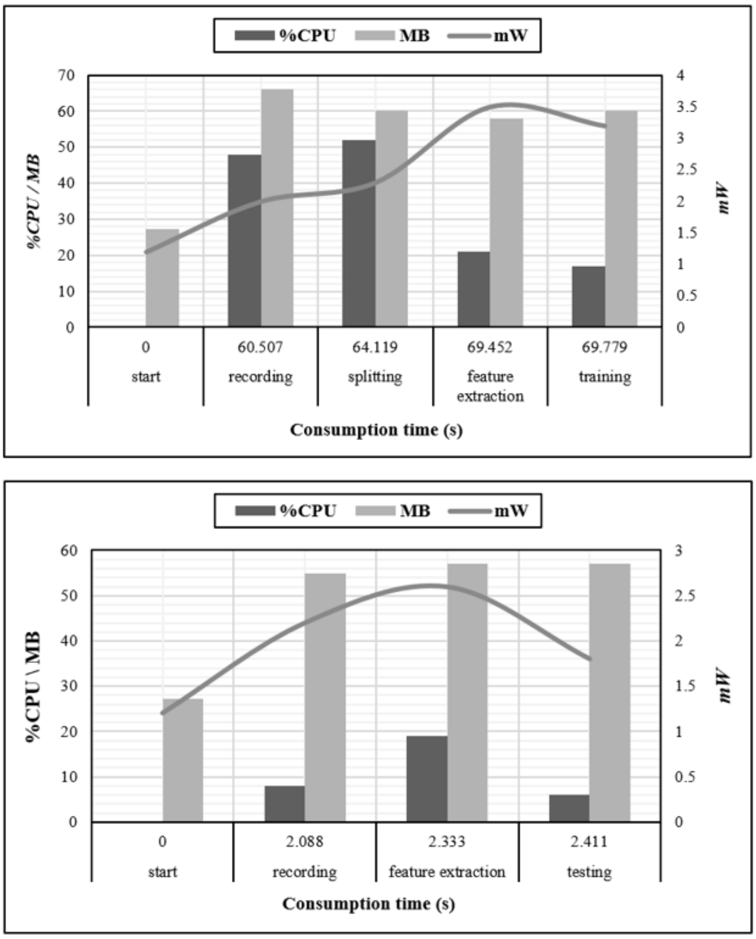


Figure 11. CPU, RAM, battery and time consumption, respectively expressed in %CPU, MB, mW and seconds, over every stage of the experiment, where the first chart refers to the training process and the second is the identification.

5.6. Participants Opinion Considerations

Here we report participants'opinions concerning the proposed system. Hence, it aims at better understanding users'needs and habits as regards authentication in order to replace present mechanisms offered on mobile devices. This survey showed that two of the five users who have enabled a knowledge-based authentication mechanism (*i.e.* PIN or pattern) have reported that it is overly repetitive and lead them to make mistakes several times a day. Besides, all three fingerprints users have mentioned a disturbing dysfunction with finger moisture. As a result, three participants over the nine who locked their device, as well as one over the two participants who did not employ such security, would use this system as a replacement of their present authentication scheme because of its simplicity. Moreover, eight respondents have mentioned they could place their confidence in the described system. However, the three remaining participants have declared that talking to their mobile device could be annoying in public areas and consequently, they have claimed that they do not trust any voice-based authentication scheme. Nevertheless, these three participants have conceded that a continuous authentication without even mind about it, was seductive and they all declared that they would be less worried to use, daily, a system that does not transmit data over the network (even if they are encrypted).

5.7. Discussion

Firstly, based on the results we have obtained in previous sections, it is possible for us to observe that the rate of correct identification remains consistent when our data set is compared to the ones we have built through both the Ted-LIUM and a subset of the TIMIT corpora. Moreover, these results stand relatively similar to the ones obtained by Kumar *et al.* [18] but not as good as the ones achieved by Nair and Salam [19] as exposed in Table 4. Nevertheless, since we have also exploited LPCCs as discriminating voice features, it is possible for us to say that our classification algorithm remains theoretically less expensive than a DTW-based solution that involves quadratic time and space complexities. Due to the use of LPCCs features, comparing our technique directly with MFCCs-based ones is very limited. However, according to the comparative detailed in Table 4 the results achieved by our proposed system also remain consistent in regard to the work of Reynolds and Rose [20]. Moreover, the reliability of our proposed system stand acceptable, in real life recording conditions, with a small number of classes; that is, the common use case of authentication mechanisms (*i.e.* the mobile device owner and potentially one or two more people). The results obtained with fraudulent utterances of speakers that participate to our experiment lead us to state that our system is perfectible in terms of fraudulent access though replayed audio samples. However the decision-making process suggested in this work should significantly reduce risks involved. Indeed, since none of the played-back samples misled the authentication mechanisms when the headset is involved, attackers will only have access to the content with a *protected* access that refers to non-critical piece of information mobile devices may contain. Hence, by introducing the notion of access privileges, we also aim at reducing unsafe situations in case of false acceptance identifications.

Secondly, the participants'opinion collection allows us to state that our system could be a relevant authentication mechanism for several users. In addition, since it is text-independent, such a system, with a few modifications, could perform the authentication in a continuous manner, without any involvement from the user. In that sense, anxieties, as regards the discomfort in talking to a device in public places, which were reported in the past may be reduced to void. Therefore, we esteem that such a technique may be a more significant option as part of a multilayer authentication. Moreover, it should also be better employed as a more reliable fallback solution in order to eradicate PIN codes.

6. Conclusions

In this research, we have proposed the design of a text-independent speaker authentication system for mobile devices with a specific focus on its usability. This implementation operates as stand-alone which does not require any network communications. Indeed, both training and identification phases, which are based on LPCCs and the Naïve Bayes classifier, are achieved onto the device itself. Moreover, we have enhanced the identification thanks to a decision-making that substantially relies on user locations and the presence of a headset. Results we have obtained over the different analysis we have performed, suggest that it embodies a reliable and efficient authentication mechanism in both quiet and noisy environments (*i.e.* 90% and 80% of kappa in quiet and noisy environments, respectively), capable of running on weakest mobile devices.

We found that 7 users were still not ready to switch from their present authentication mechanism. Moreover, three of the participants have reported that they could not place their confidence in such a system, as it may be disturbing when used in public places. However, since it is text-independent, legitimate users may be implicitly authenticated as they start speaking, insofar as the mobile device is neither in their pocket, nor their bag (*i.e.* during a conversation). In that sense, since the idea of being authenticated in a continuous manner was seducing to skeptical participants, we also suggest that this technique should be either used in a multilayer authentication system, or as a fallback mechanism, namely when the first one fails, to cover most of the users'needs and usages.

7. Future works

Future works will focus on offering the application on the *Google Play Store* to better assess the accuracy and the robustness of the proposed authentication system. However, the current implementation will be adapted in order to let us track user authentication attempt outcomes and locations. In this way, such a large-scale evaluation will provide more reliable results in front of real life condition usages and the location-based decision will be better exploited and significant, as it was in the experiment we have conducted in this research. Besides, the extraction of MFCCs discriminating voice features will be considered in order to produce a direct comparison in terms of reliability and effectiveness with LPCCs features.

Acknowledgments: Authors would like to acknowledge Sébastien Gaboury and Valère Plantevin for their conscientious reviews, as well as every person who participates in our experiment.

Author Contributions: The authors contributed in equal parts.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
CDBN	Convolutional Deep Belief Networks
DBFS	Decibels Relative to Full Scale
DTW	Dynamic Time Warping
EER	Equal Rerror Rate
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LPC	Linear Prediction Coefficient
LPCC	Linear Prediction Cepstral Coefficient
MFCC	Mel-Frequency Cepstral Coefficient
PCM	Pulse-Code Modulation
PIN	Personal Identification Number
SPH	NIST Sphere Format
VAD	Voice Activity Detection
VQ	Vector Quantization
WAV	Waveform Audio File Format

References

1. Laurence, G.; Janessa, R. Market Share: Devices, All Countries, 4Q14 Update. Report, 2015.
2. Wilska, T.A. Mobile phone use as part of young people's consumption styles. *Journal of consumer policy* **2003**, *26*, 441–463.
3. Goggin, G. *Cell phone culture: Mobile technology in everyday life*; Routledge, 2012.
4. Falaki, H.; Mahajan, R.; Kandula, S.; Lymberopoulos, D.; Govindan, R.; Estrin, D. Diversity in smartphone usage. Proceedings of the 8th international conference on Mobile systems, applications, and services. ACM, 2010, pp. 179–194.
5. Ben-Asher, N.; Kirschnick, N.; Sieger, H.; Meyer, J.; Ben-Oved, A.; Möller, S. On the need for different security methods on mobile phones. Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM, 2011, pp. 465–473.
6. Yan, J.; Blackwell, A.; Anderson, R.; Grant, A. Password memorability and security: Empirical results. *IEEE Security & privacy* **2004**, *2*, 25–31.
7. Clarke, N.L.; Furnell, S.M. Authentication of users on mobile telephones–A survey of attitudes and practices. *Computers & Security* **2005**, *24*, 519–527.
8. Clarke, N.L.; Furnell, S.M.; Rodwell, P.M.; Reynolds, P.L. Acceptance of subscriber authentication methods for mobile telephony devices. *Computers & Security* **2002**, *21*, 220–228.

9. Yampolskiy, R.V. Analyzing user password selection behavior for reduction of password space. Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International. IEEE, 2006, pp. 109–115.
10. Bond, R.H.; Kramer, A.; Gozzini, G. Molded fingerprint sensor structure with indicia regions, 2012. US Patent D652,332.
11. Derawi, M.O.; Nickel, C.; Bours, P.; Busch, C. Unobtrusive user-authentication on mobile phones using biometric gait recognition. Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). IEEE, 2010, pp. 306–311.
12. Gafurov, D.; Helkala, K.; Søndrol, T. Biometric Gait Authentication Using Accelerometer Sensor. *JCP* **2006**, *1*, 51–59.
13. Holz, C.; Buthpitiya, S.; Knaust, M. Bodyprint: Biometric User Identification on Mobile Devices Using the Capacitive Touchscreen to Scan Body Parts. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 3011–3014.
14. Eriksson, A.; Wretling, P. HOW FLEXIBLE IS THE HUMAN VOICE?–A CASE STUDY OF MIMICRY. *Target* **1997**, *30*, 29–90.
15. Doddington, G.R.; Przybocki, M.A.; Martin, A.F.; Reynolds, D.A. The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech Communication* **2000**, *31*, 225–254.
16. Gold, B.; Morgan, N.; Ellis, D. *Speech and audio signal processing: processing and perception of speech and music*; John Wiley & Sons, 2011.
17. Jain, A.; Bolle, R.; Pankanti, S. *Biometrics: personal identification in networked society*; Vol. 479, Springer Science & Business Media, 2006.
18. Kumar, R.; Ranjan, R.; Singh, S.K.; Kala, R.; Shukla, A.; Tiwari, R. Multilingual speaker recognition using neural network. Proceedings of the Frontiers of Research on Speech and Music (FRSM-2009), 2009, pp. 1–8.
19. Nair, R.; Salam, N. A reliable speaker verification system based on LPCC and DTW. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2014, pp. 1–4.
20. Reynolds, D.A.; Rose, R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing* **1995**, *3*, 72–83.
21. Thullier, F.; Bouchard, B.; Ménélas, B.A.J. Exploring Mobile Authentication Mechanisms from Personal Identification Numbers to Biometrics Including the Future Trend. *Protecting Mobile Networks and Devices: Challenges and Solutions* **2016**, p. 1.
22. Milanesi, C. Voice Assistant Anyone? Yes please, but not in public!, 2016.
23. Rabiner, L.R.; Juang, B.H. Fundamentals of speech recognition **1993**.
24. LeCun, Y.A.; Bottou, L.; Orr, G.B.; Müller, K.R. Efficient backprop. In *Neural networks: Tricks of the trade*; Springer, 2012; pp. 9–48.
25. Salvador, S.; Chan, P. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis* **2007**, *11*, 561–580.
26. Lee, H.; Pham, P.; Largman, Y.; Ng, A.Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in neural information processing systems, 2009, pp. 1096–1104.
27. Reynolds, D.A. Speaker identification and verification using Gaussian mixture speaker models. *Speech communication* **1995**, *17*, 91–108.
28. Plieninger, A. Deep Learning Neural Networks on Mobile Platforms **2016**.
29. Movidius. Google and Movidius to Enhance Deep Learning Capabilities in Next-Gen Devices, 2016.
30. Rao, K.S.; Vuppala, A.K.; Chakrabarti, S.; Dutta, L. Robust speaker recognition on mobile devices. 2010 International Conference on Signal Processing and Communications (SPCOM). IEEE, 2010, pp. 1–5.
31. Brunet, K.; Taam, K.; Cherrier, E.; Faye, N.; Rosenberger, C. Speaker Recognition for Mobile User Authentication: An Android Solution. 8ème Conférence sur la Sécurité des Architectures Réseaux et Systèmes d'Information (SAR SSI), 2013, p. 10.
32. Lindberg, J.; Blomberg, M.; others. Vulnerability in speaker verification-a study of technical impostor techniques. Eurospeech, 1999, Vol. 99, pp. 1211–1214.
33. Sadjadi, S.O.; Hansen, J.H. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Processing Letters* **2013**, *20*, 197–200.

34. Benesty, J.; Sondhi, M.M.; Huang, Y. *Springer handbook of speech processing*; Springer Science & Business Media, 2007.
35. Al-Hassani, M.D.; Kadhim, A.A. Design a text-prompt speaker recognition system using LPC-derived features. The 13th International Arab Conference on Information Technology ACIT, 2012, pp. 10–13.
36. Kinnunen, T. Spectral features for automatic text-independent speaker recognition. *Licentiate's Thesis, University of Joensuu.–2003* **2003**.
37. John, G.H.; Langley, P. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.
38. Zhang, H. The optimality of naive Bayes. *AA* **2004**, *1*, 3.
39. Rousseau, A.; Deléglise, P.; Estève, Y. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. LREC, 2014, pp. 3935–3939.
40. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n* **1993**, *93*.
41. Ben-David, A. A lot of randomness is hiding in accuracy. *Engineering Applications of Artificial Intelligence* **2007**, *20*, 875–885.
42. Villalba, J.; Lleida, E. Preventing replay attacks on speaker verification systems. IEEE International Carnahan Conference on Security Technology (ICCST), 2011, pp. 1–8.