cryptography

MDPI

Article

A Text-Independent Speaker Authentication System for Mobile Devices

Florentin Thullier *, Bruno Bouchard * and Bob-Antoine J. Menelas *

Department of Computer Science and Mathematics, Université du Québec à Chicoutimi, Chicoutimi, QC G7H 2B1, Canada

* Correspondence: florentin.thullier1@uqac.ca (F.T.); bruno_bouchard@uqac.ca (B.B.); bob-antoine-jerry_menelas@uqac.ca (B.-A.J.M.)

Received: 6 July 2017; Accepted: 19 September 2017; Published: 22 September 2017

Abstract: This paper presents a text independent speaker authentication method adapted to mobile devices. Special attention was placed on delivering a fully operational application, which admits a sufficient reliability level and 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 on the mobile device through the extraction of linear prediction cepstral coefficients and the naive Bayes algorithm as the 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 quiet and noisy environments. Public speech corpora were also employed to compare this implementation to existing methods. Results were efficient regarding mobile resources' consumption. The overall classification performance obtained was accurate with a small number of samples. Then, it appeared that our authentication system might be used as a first security layer, but also as part of a multilayer authentication, or as a fall-back mechanism.

Keywords: speaker authentication; text independent; mobile devices; LPCCs; naive Bayes; voice; security

1. Introduction

Nowadays, mobile devices play an important role in humans'activities. As announced by the Gartner Institute, smartphone sales surpassed one billion units in 2014 [1], and everywhere and at all times, people carry their mobile devices [2] (considered as a vital piece of their life) [3]. People store private information, like pictures and recordings, and also secret data (i.e., emails, bank account) on their devices. However, they generally do not really pay attention regarding the safety of this secret content [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.

In the last decade, multiple authentication methods have been designed and evaluated. Mobile devices often only offer authentications that involve recalling a piece of information such as the PIN code. However, they concede several drawbacks. For instance, it was accounted that half of the population leaves the mobile devices unlocked [5]. They evaluate that entering a PIN code includes loads of burden for each time the cell phone must be opened [5]. Besides, it is realized that users experience difficulties recalling all passwords that they utilize these days [6]. Unmistakably, these practices may lead to a tremendous effect on the security of cell phones. Doing so, people's authentication usage may create important dangers for the security that a method offers at first [7–9]. Recently, biometric authentication mechanisms such as fingerprint, ear shape or gait recognition were enabled on mobile devices [10–13]. These systems chiefly exploit the uniqueness

of the user's physiological or behavioural 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 concentrate on both vocal qualities delivered by the discourse and the discourse itself. These components rely upon the measurement of the vocal tract, mouth and nasal depressions, yet additionally depend on voice pitch, talking style and dialect [14].

Speaker validation frameworks might be composed by two driving strategies: text-dependent and text independent [15,16]. In a text-dependent authentication, the person has to say a predefined pass-phrase, seen as a voice secret word. This predefined phrase is utilized both for the enrolment and the recognizable proof process. For example, on the Android mobile operating system, the expression "Ok Google" is exploited as a predefined pass-expression. It has to be vocalized for both enrolment and identification. By contrast, text independent methods are capable of authenticating the speaker regardless of the pronounced expression. Nowadays, speaker verification strategies offered on mobile devices deeply rely on matching templates methods realized through cloud computing. Because of that, additional costs may be associated. Nevertheless, such solutions can still be considered as being inexpensive considering that they do not necessitate any extra sensors. Conversely, since manufacturers have pushed fingerprint systems to the forefront of the mobile device 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 provide an adequate acceptance percentage. They seem to be less invasive than fingerprint or retina scan [8,17]. Moreover, these approaches may assume a noteworthy part in everyday life as some applications, like e-commerce solutions, attendance systems, mobile banking or forensics, need to be secured.

Experiments have proven that some proposed speaker recognition and identification systems achieve accurate results [18–22]. In spite of their effectiveness, few of such mechanisms have been exploited in real life. They are mostly machine-centred implementations. Additionally, the significant number of users who still do not secure the entrance to their cell phones [5] reveals a need for novel methods mainly focused on a human-centred design that must take into account the diversity of user profiles and usages [23]. The current initiative addresses these observations. The contribution of this paper is to expose the design of a Text independent Speaker Authentication (TiSA) system suitable for mobile devices while focusing on users'needs. The choice 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 of identifying and authenticating the owner of the mobile device all along the conversation without any care for what is being said. Moreover, a recent study [24] 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 [25]. 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 naive Bayes algorithm for patterns' classification. Whereas authentication methods usually either grant or deny access to the whole content of the phone, a privileged access is also introduced, in this work, to be able to face false positive and negative authentications. Here, some accesses, based on a simple evaluation of the user's location and the presence of a headset, may be granted or rejected. To produce an efficient system, 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 contribution of the paper is the following. Section 2 briefly reviews the literature about speaker identification and verification systems. The third one describes the proposed approach. Section 4 describes the experiments we conducted in order to evaluate the reliability, as well as the efficiency of

such an implementation. Section 5 exposes and discusses the results we obtained. Finally, Section 6 draws the conclusion, and Section 7 provides future works.

2. Related Work

In the past few years, multiple algorithms have been developed to authenticate a speaking person. Most of these algorithms focused on the extraction of features and classifications. In this section, we will briefly review proposed techniques, as well as the evaluation of their adaptability regarding the mobile context. We will also review speaker authentication techniques designed to be run on mobile devices.

One of the first works of this category was proposed by Reynolds and Rose [20]. The authors suggested the use of Mel-Frequency Cepstral Coefficients (MFCCs) as features and a Gaussian Mixture Model (GMM) to authenticate the speaker. MFCCs characterize adequately the envelope of the short-time power spectrum of the signal. In spite of being sufficiently resilient to noisy conditions, their applicability in the mobile context is limited as they monopolize many resources [26]. The argumentation that supports this method comes from an experiment with a subset of the "KING speech" [27] database. This database provides utterances from speaker conversations over both signal-to-noise radio channels and narrow-band telephone channels. It has been observed that a large number of unlabelled classes of the sample distribution may be encoded as a linear combination of Gaussian basis functions. An accuracy of 80.8% was obtained for 49 telephone speech samples of 15 s. The authors hypothesized that this model should be computationally inexpensive and easy to implement on real-time platforms. However, the main drawback of their method comes from the initialization of the training process as several parameters such as the mean, covariance and prior of each distribution have to fit the data. 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 would probably be computationally overly expensive.

A second text independent method for speaker authentication has been described in [18]. This methods relies on a back-propagation neural network having LPC (Linear Prediction Coefficient) parameters as input features, to predict utterances. The use of the back-propagation method aims to optimize the distribution of weights between neuron layers. Doing so, it becomes possible for the neural network to correctly map arbitrary inputs to outputs. The decision process is achieved by associating each speaker to an utterance. In a database having 25 speech samples of different languages, the identification accuracy was about 85.74%. With this promising achievement, Kumar et al. [18] have concluded that the proposed method would be appropriate and reliable. Nevertheless, we may note that the theoretical complexity of a standard back-propagation neural network training phase is $O(nmh^k oi)$. Here, 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 [28]. This suggests that the computation time is still overly expensive considering the limited capacity of mobile devices.

Another text independent method for speaker authentication has been proposed by Nair and Salam [19]. This method resorts to both LPCs and LPCCs to compare their strength. The use of the Dynamic Time Warping (DTW) [29] algorithm allows to decide about the best option. The TIMIT (Texas Instruments and Massachusetts Institute of Technology) speech database was used for the experiment. This corpus of American-English Speakers (AESs) counts 630 speech signals. The achieved accuracy is around 92.2% with LPCs. With derivative cepstral coefficients, it climbed to 97.3%. As expected, the association of LPCCs to the DTW algorithm offers an accurate and reliable solution. Since DTW requires a quadratic complexity both in terms of time and memory usage (i.e., $O(n^2)$) [30], it appears that it may not be the most suitable solution to achieve speaker authentication, directly on the mobile device. Nevertheless, real speaker authentication scenarios usually imply few distinct samples. In that sense, DTW for decision-making still remains an acceptable choice for such an authentication mechanism on limited-performance devices especially when considering other fields of research such as in [31,32].

The growing interest in deep leaning approaches observed in recent years forced us to question its suitability as regards a speaker identification task. Lee et al. [33] have shown that Convolutional Deep Belief Network (CDBN) features trained with an SVM classifier have outperformed MFCC features trained with a GMM (respectively for the method described in [34] when the number of training examples was small (i.e., 1–2 utterances per speaker). However, with a greater number of training samples (i.e., 5–8 utterances per speaker), the 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 [35]. 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 [36]. Then, in the present situation, it appears that such an approach may not be an adequate solution according to our needs.

To the best of our knowledge, it seems that a limited number of text independent speaker authentication methods have been implemented on mobile devices. For instance, to tackle TiSA solutions in noisy conditions, Vuppala et al. [37] suggested a recognition method that exploits various speech enhancements in order to enhance overall performances. However, realized evaluations, with the TIMIT database, were performed with various simulated noises.

On the other hand, Brunet et al. have proposed in [38] a TiSA method dedicated to mobile devices. This method starts by extracting MFCC features from speech samples. With this information, a Vector Quantization (VQ) method allows to construct a reference model. Euclidean distances between stored centroids and tested samples enable to accept or to reject the attempt based on a given threshold. To evaluate the proposed method, the Sphinx dataset [39] that counts 16 utterances of AESs, as well as a personal database are exploited. For their personal database, samples for training and testing were collected with a mobile device. Being implemented as a stand-alone biometric system, the Equal Error Rate (EER) was the only performance indicator. 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). As usually observed for such an approach, the observed results are deeply dependent on the initial parameters, like the number of centroids.

With this short analysis of the literature, we observe that few TiSA methods have considered the limited capabilities of mobile devices. This paper introduces a user-centred TiSA system for mobile devices. 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 other expensive techniques regarding processing time, as long as they offer an accurate identification.

3. Proposed Speaker Authentication System

In this paper, we propose a new authentication system for mobile devices based on speakers, which is text independent. This system is fully stand-alone because all the processing is done directly on the mobile device. Therefore, our approach does not require any heavy (and costly) client/server platform. More precisely, the architecture of our system is composed of three fundamental processes, as you can see in Figure 1. The first process consists of extracting a selected set of individual voice features from an audio signal coming from the speaker, in order to build a dataset. The second process takes that dataset as input and performs a training exploiting a naive Bayes classifier. Finally, the third process computes the authentication decision and returns true or false.

The objective of our system is to enhance the standard speaker verification mechanism in order to increase the confidence rate. To do that, we propose to grant specific access privileges to the user by evaluating two discriminant variables. The first variable is the actual location of the person versus the one defined beforehand. The second variable is the presence (or not) of the headset, which consists simply of checking if it is plugged into the mobile device. Of course, using a headset with a built-in

microphone proceeding to a noise reduction will decrease the possibility of a user being unwillingly authenticated though replay attacks [40].

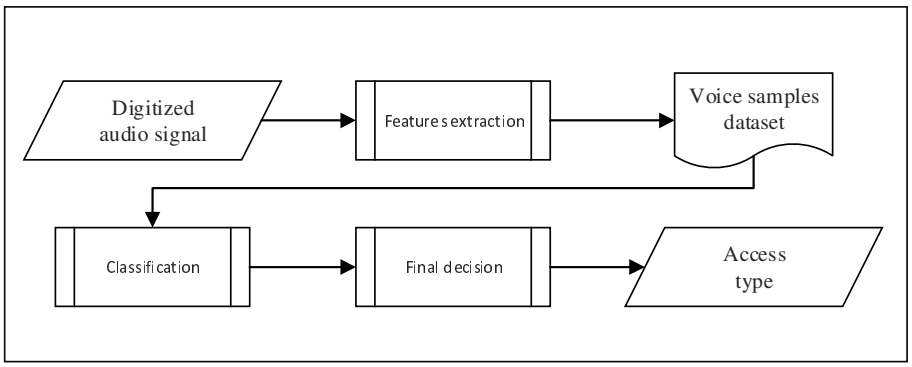

Figure 1. Flowchart of our proposed speaker authentication system.

3.1. Input

Each audio file is recorded with a 16-bit signed integer PCM encoding format using bi-channels. The sampling rate of this kind of audio file is fixed to 44.1 kHz.

3.2. Pre-Processing

This section will describe the pre-processing phase of our approach. This phase consists of two main steps, which are voice activity detection and audio normalization.

3.2.1. Voice Activity Detection

The first step of the pre-processing phase consists of trimming the audio file to remove every silence interval in order to keep only speech segments. To achieve that, we defined a fixed threshold close to zero (i.e., 0.0001). We use that threshold to identify the sections of the input signal that we need to remove (i.e., the ones that are close to it). Then, we apply the autocorrelation function $r_x(t,k)$ introduced by Sadjadi and Hansen [41] onto a windowed audio segment $s_w(n)$ of the complete input signal $s(n)$ given by,

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

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

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

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

As we can see on Figure 2, for each processed segment $s_w(n)$, if the mean value of the computed coefficients that result from the autocorrelation function gets close to the fixed threshold, then the segment is identified as a silence interval, and thus, it is removed.

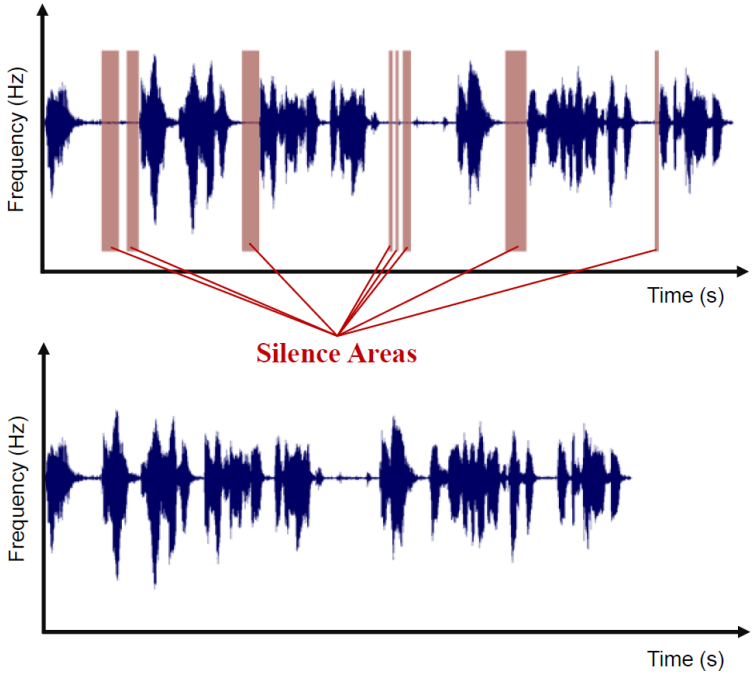

Figure 2. The first signal is the raw input where silence areas are highlighted. The second is the output of the same signal after the silence removal process.

3.2.2. Audio Normalization

After the silence removal phase, we perform a peak normalization. The objective is to modify the gain of the input to the highest peak of the signal, uniformly. Normally, this process allows ensuring that the highest peak remains at zero decibels relative to Full Scale (dBFS), which is the loudest level allowed in a digital system. It should be noted that the entire signal is adjusted so the original information is not affected. In addition, peak normalization ensures that the audio signal will not clip in any ways. The result of this process is shown in Figure 3.

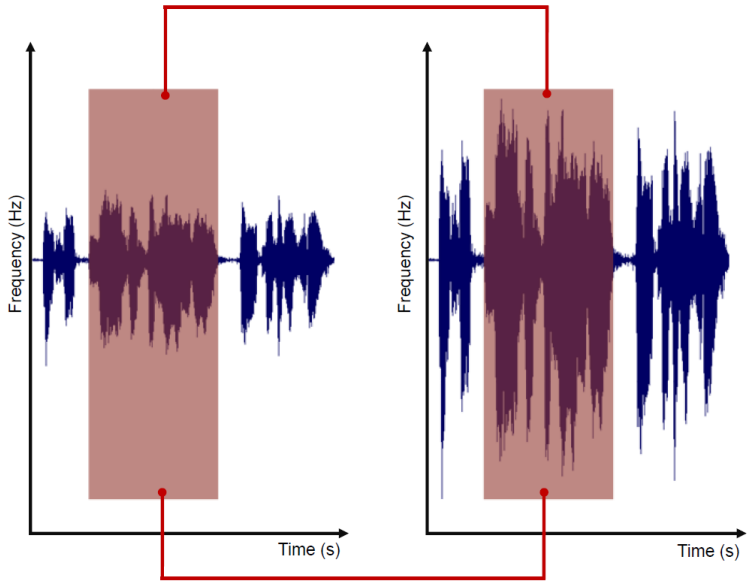

Figure 3. The left signal is the input signal, and the right one is the same signal with peak normalization, where the same sequence is highlighted on both signals.

3.3. Feature Extraction

Extracting discriminating features from an audio signal containing the voice of a speaker is not trivial. The voice is considered as a very particular signal containing rich information about the speaker. Therefore, extracting features from the speech constitutes a core component of both speaker identification and authentication systems. In our approach, we propose to exploit Linear Prediction Cepstral Coefficients (LPCCs) to perform the features extraction. Such coefficients are directly derived from the linear prediction analysis, which aims to estimate the relevant characteristics from a speech signal [42]. Using LPCCs allows us to provide very accurate estimates of the speech parameters while keeping a good computation speed [26]. Mobile devices have limited computational resources compared to a standard computer; thus, choosing a method that required low computational power is essential. Each step from a pre-processed signal to the obtaining voice characteristics that are saved in a dataset are summarized in Figure 4.

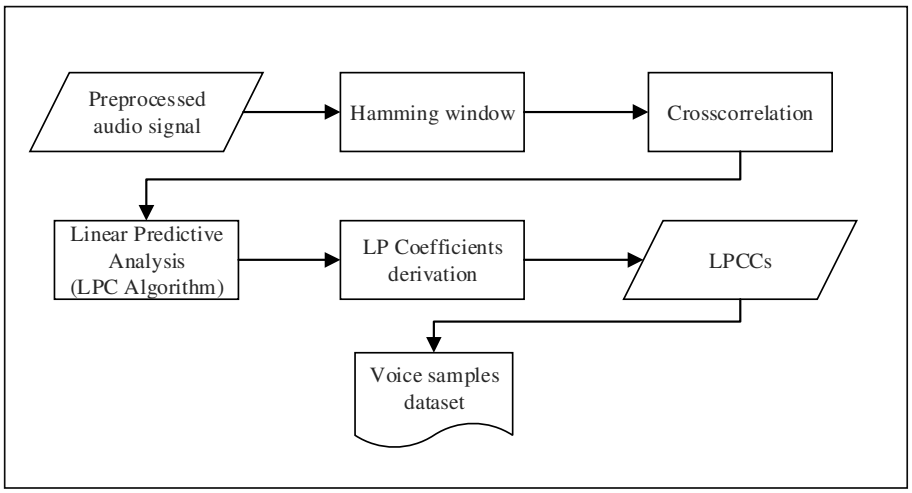

Figure 4. Flowchart of the feature extraction process.

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 25 ms 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 [43]. 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. 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 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 matrix form as,

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

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 equal.

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 [44], we suggest using 20 LPCCs to preserve a relatively good computation speed.

3.4. Classification

In the literature, several classification algorithms have been used for speaker recognition (i.e., GMM, ANN, etc.). Nevertheless, in our context, we needed an algorithmic approach requiring

low computational resources. It is well known that naive Bayes classifiers are fast, very effective and easy to implement. This method consists of a supervised and statistical learning algorithm for classification, which computes the conditional probabilities of the different classes given the value of attributes. At the end, it selects the class with the highest conditional probability. Table 1 shows the formal complexity evaluations (time and space) of the naive Bayes classifier [45].

Table 1. Naive Bayes time and space complexities, given k features for both training and testing operations [45].

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

More precisely, the method works as follows (see Figure 5). 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 computed. Each sample has k features (i.e., LPCCs) represented by floating numbers (with $k = 20$), which are denoted as $a_1, a_2, \ldots, a_n$. The goal of the classifier is to use these samples to build a model (in the training phase) that will be exploited 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.

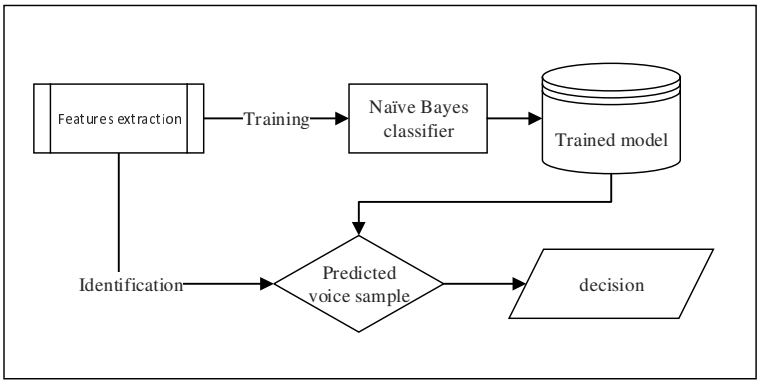

Figure 5. Flowchart of the classification process.

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 naive Bayes classifier. Nevertheless, Zhang [46] has demonstrated that such a condition is not necessary to satisfy in practical situations. Indeed, no matter how strong dependencies among attributes are,

naive Bayes can still be optimal if these 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 naive Bayes according to the supposed quality of the LPCCs.

3.5. Decision-Making

The decision-making process (granting access or not) is a crucial phase in our system. This process is illustrated in detail in Figure 6. It may lead to two kinds of errors. First, it may result in a false negative, which means that the system fails to identify a genuine user. Secondly, it may result in a false positive, which means granting access to a non-authorized user. While a false negative authentication does not compromise the security or the privacy of the user's data, it constitutes a huge source of frustration. In that case (false negative), the authentication process has to be redone, or a fall-back mechanism (i.e., a PIN number) must be used. On the other hand, a false positive authentication poses a serious vulnerability issue. Besides, speaker authentication systems have limitations that may also lead to security threats. For instance, they are vulnerable to voice imitation or false authentication exploiting legitimate voice records.

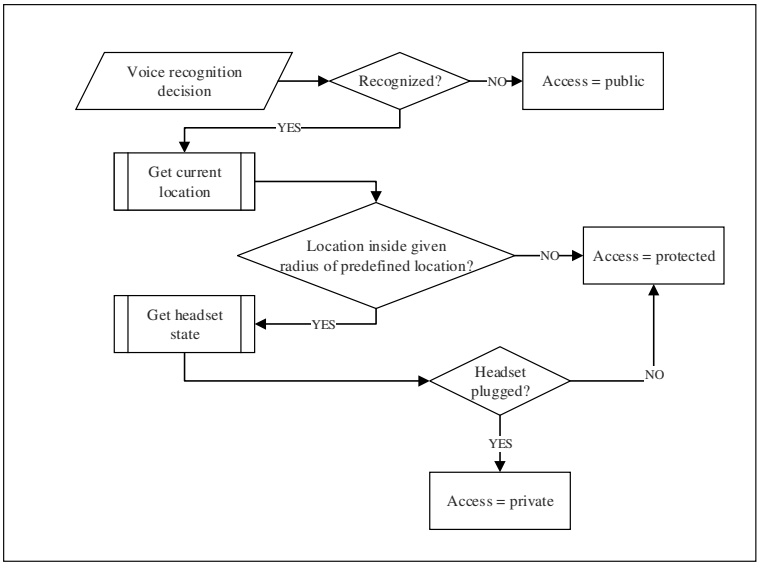

Figure 6. Flowchart of the decision-making process.

To address these vulnerability issues, we propose to introduce the notion of access privileges to prevent against misidentification. First, the user needs to select the privilege preferences for each application on his/her mobile device. Three types of setting may be selected: public, protected or private. Of course, a user logged with public privileges will only be allowed access to non-critical content and applications. A user logged with protected privileges will have access to most of the content, but with some restriction on sensitive data or applications (i.e., bank account). Finally, a user logged with private privileges will have access to all the content of the mobile device.

Now, the question is: how to choose which level of privileges to grant to the user based on his/her voice identification. Our system selects the safest level to grant by evaluating the result of the identification process. If the voice does not match at all with any users of the device, the system allows a public access. In the case of a false negative, the user will have to repeat the authentication process to get better privileges, but he/she still can use the public applications. If the voice authentication process finds a match in the dataset, a protected access is granted, and then, the current location of the device is fetched. Thereafter, the system check if the device is actually in what we call "a trusted location". These trusted locations are predefined by the user (i.e., home, work, etc.). The acceptable radius of

the location may be adjusted, but for our prototype, we used a radius of 200 and 500 m. This double validation (i.e., voice and location) allows the system to be much more robust against fraudulent authentication attempts. Indeed, the system is not perfect, and there are still risks. For instance, people from the same family or co-workers are well aware and often share most of the usual locations of the user. To minimize these risks, we offer yet another verification process. In order to be granted a private access level, the user must use a headset to identify himself/herself to the device, and all previous verification must be satisfied. Thus, the system checks if the headset is plugged into the device, providing extra microphones that enhance the accuracy of the voice recognition. These microphones are closer to the mouth and therefore they provide better noise filtering than the built-in microphone of the mobile device. By adopting this strategy, we think that the false positive rate will considerably decreased. Of course, the proposed approach is experimental, but it is easy to foresee how it could be standardized for all mobile devices.

We believe that all three suggested privacy options will allow a certain flexibility in the access of the mobile device. Indeed, we provide a means to define an appropriate level of restriction since each user has many different considerations according to what piece of information he/she has on his/her phone that is important as regards confidentiality or not. Users are thus able to tune the system to make the decision-making more or less restrictive and to best fit their personal needs.

4. Experiments

The authentication system proposed in this paper has been tested with 11 speakers, which were students recruited at our university. We designed an experiment aiming to assess the effectiveness of the system. In this experiment, each participant was asked to use the developed system for authenticate himself/herself on a provided device. Each device was equipped with a headset plugged. We tested the system in two different environmental contexts. First, the training phase was completed, for each participant, in a quiet environment, limiting the noise. Thereafter, an authentication attempt has been tried in the same quiet environment. In order to test the robustness of the system, a second authentication attempt was performed in a noisy environment.

4.1. Participants

For the experiment, 11 students have been recruited, including seven males and four females, aged from 19–36 years. All participants were native French speakers. However, some participants had distinctive accents, such as Canadian French and Hexagonal French. All of them were familiar with iOS and/or Android and owned at least one recent mobile device (i.e., smartphone or tablet). Finally, nine of the participants used, on a regular basis, an unlocking mechanism for their smart device (PIN: 4, pattern: 2, fingerprint: 3), and fingerprint users either had a PIN code or a pattern as fall-back mechanism.

4.2. Data Collection

The prototype of our text independent system (Figure 7 shows a screenshot) was implemented on an Android platform as an independent application requiring a 4.0.1 version (or higher) of the mobile operating system. Each volunteer carried out the experiment using the same smartphone. The chosen model was a LG Nexus 5 running Android 6.0.1 with a Snapdragon 800 Quad-core at 2.3 GHz CPU and 2 GB of RAM. They also used the same headset (i.e., Bose SoundTrue I), and they performed the experiment in the same conditions (i.e., a room and a public place).

For the first part of the experiment, a quiet room (i.e., meeting room) was selected to perform the training and the first attempt of identification in a quiet environment. Thereafter, the participants were asked to move to the university's cafeteria at a proper time to proceed to the second attempt in a noisy environment. We measured the level of sound in each environment before the test session in order to control the conditions. For that, we exploited a sound level meter embedded in the application.

In the quiet room, the average level of sound was 16.5 dB. In the cafeteria (i.e., the noisy environment), the sound level reached 95 dB.

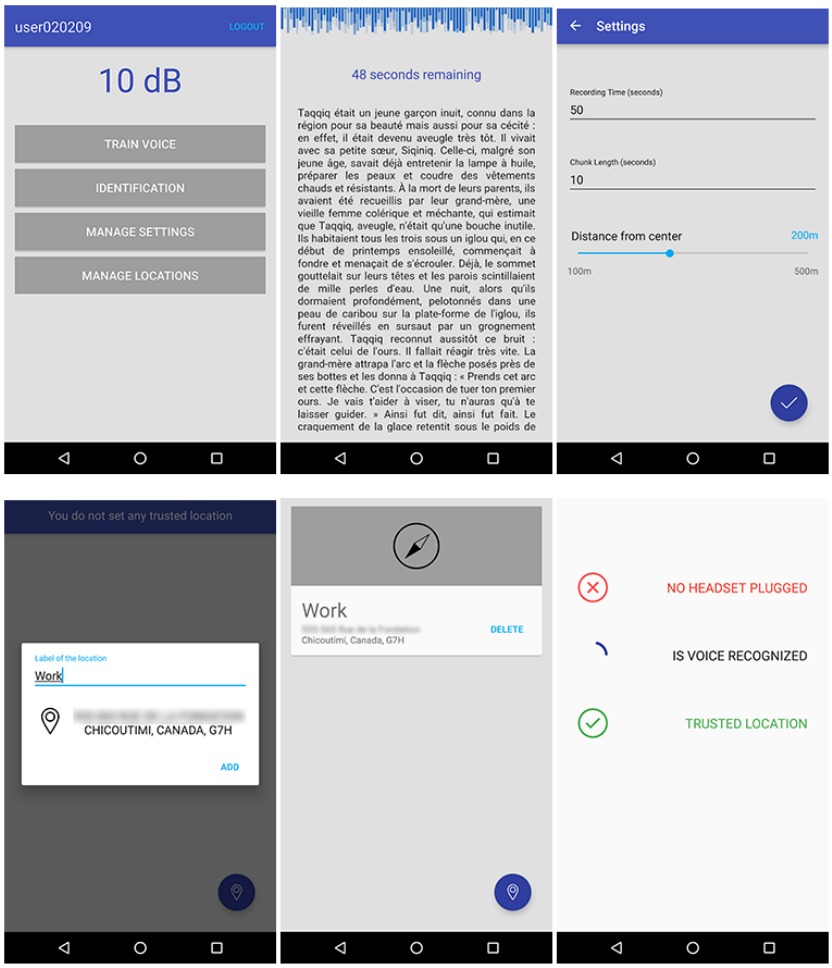

Figure 7. Screen captures of the Android application.

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 were 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 themselves 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-s chunks, 6 instances per class in total. Each set of features from each instance was written in the dataset, which was used to create the training model of the naive 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 that 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 attempt was performed in the same place, our decision-making has always stated that users stood in a trusted location. Therefore, we have mocked a location that 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 given feedback about their habits concerning authentication on their own device, as well as their opinion as regards the proposed system.

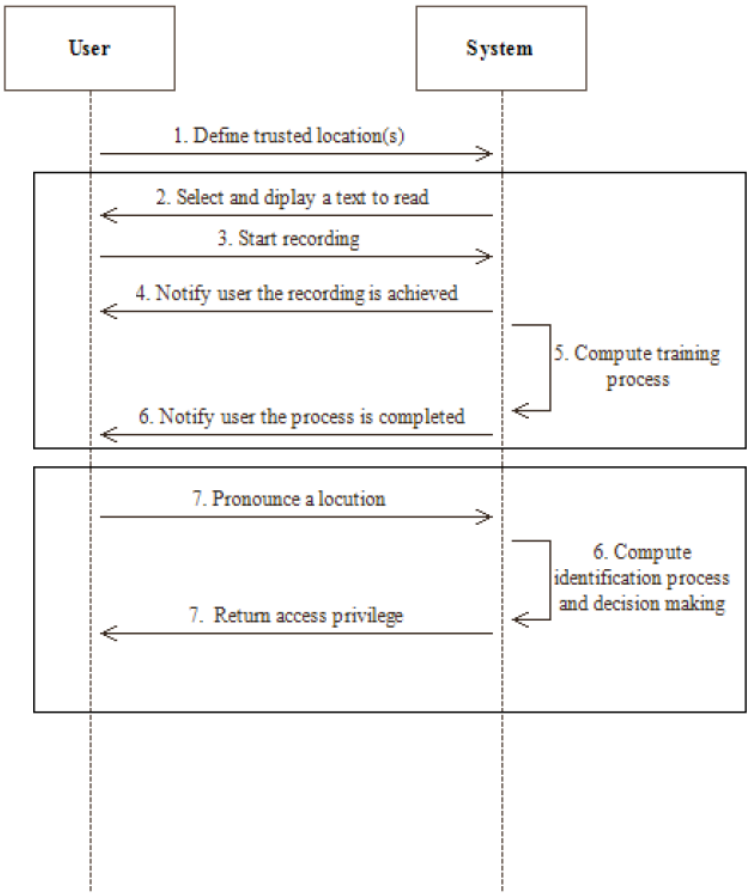

Figure 8. Sequence diagram of the experiment.

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 purposes with the dataset we suggested. The first one is the Ted-LIUM (*Laboratoire d'Informatique de l'Université du Mans*) corpus, which was proposed by Rousseau et al. [47]. 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 converting the whole files into Waveform Audio File Format (WAV). Furthermore, we took care of removing the first fourth frame 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. [48]. 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 to which registered speaker a given utterance corresponds, 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 includes false positive (FP) and false negative (FN) predictions.

Despite its popularity, accuracy alone does not typically 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 [49], we decided to provide 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, the results of the experiment that 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. Our dataset was named UQAC-Studs that stands for students of the Université du Québec à Chicoutimi.

Table 2. Results of the experiment based on the realized dataset: 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 datasets 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 six samples to construct the training set, and the four remaining were used for the identification. The results we obtained over this speech corpus are shown in Table 3.

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

TIMIT Subset (16 Speakers)	
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 datasets 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 s. In order to be more consistent with our experimental procedure, the first six 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%.

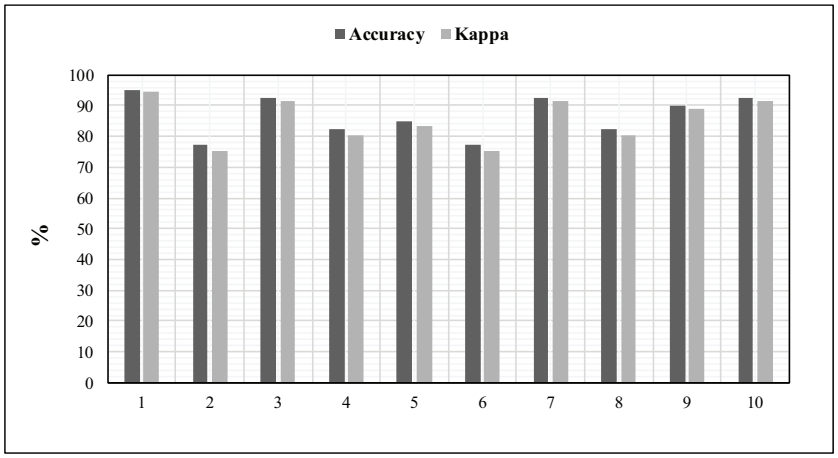

Figure 9. Accuracy and kappa measures achieved by our system over the 10 random batches of the Ted-LIUM corpus we have created.

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.

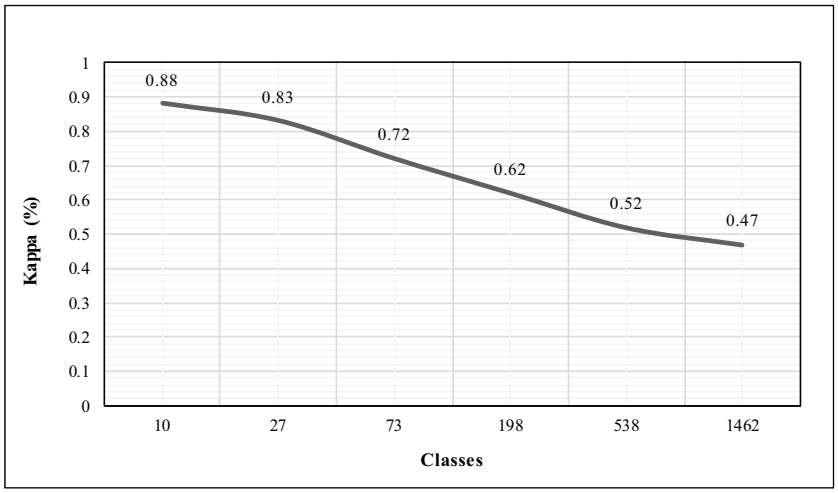

Figure 10. Evolution of the kappa measure over the Ted-LIUM corpus when increasing the number of classes exponentially.

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

Table 4. Empirical comparison between our TiSA method and previous works.

	Features	Number of Features	Classification, Pattern Matching (Training Complexity)	Accuracy	Dataset	Number of Samples
Suggested method	LPCCs	20	Naive Bayes (O_{nk})	91∼82% 87% (AVG) 83%	UQAC-Studs Ted-LIUM TIMIT (subset)	11 1462 16
Nair and Salam [19]	LPCs and LPCCs	20, 30 and 40	DTW (O_{n^2})	90.4% (20 LPCs) 94.8% (20 LPCCs)	TIMIT	630
Reynolds and Rose [20]	MFCCs	100 12-dimensional vectors per second	GMM (may vary between implementations to fit the Gaussian model)	96.8% 80.8%	KING Private samples	49
Kumar et al. [18]	LPCs, LPCCs, RC, LAR, ARCSIN and LSF	N.A.	ANN with backpropagation (O_{nmlk^iOi})	85.74%	Private samples	25

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 [40]. 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 the robustness of our system as it stands. Indeed, no specific method to counteract replay attacks such as [50] has been implemented in this work.

In order to proceed with 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, no fraudulent samples were correctly identified while using the headset, which has its own microphone embedded.

5.5. Computation Performances Considerations

Since we desired to create a user-centred TiSA 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 at every stage of the training, as well as the identification processes, for one given instance of the dataset 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 s. 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.

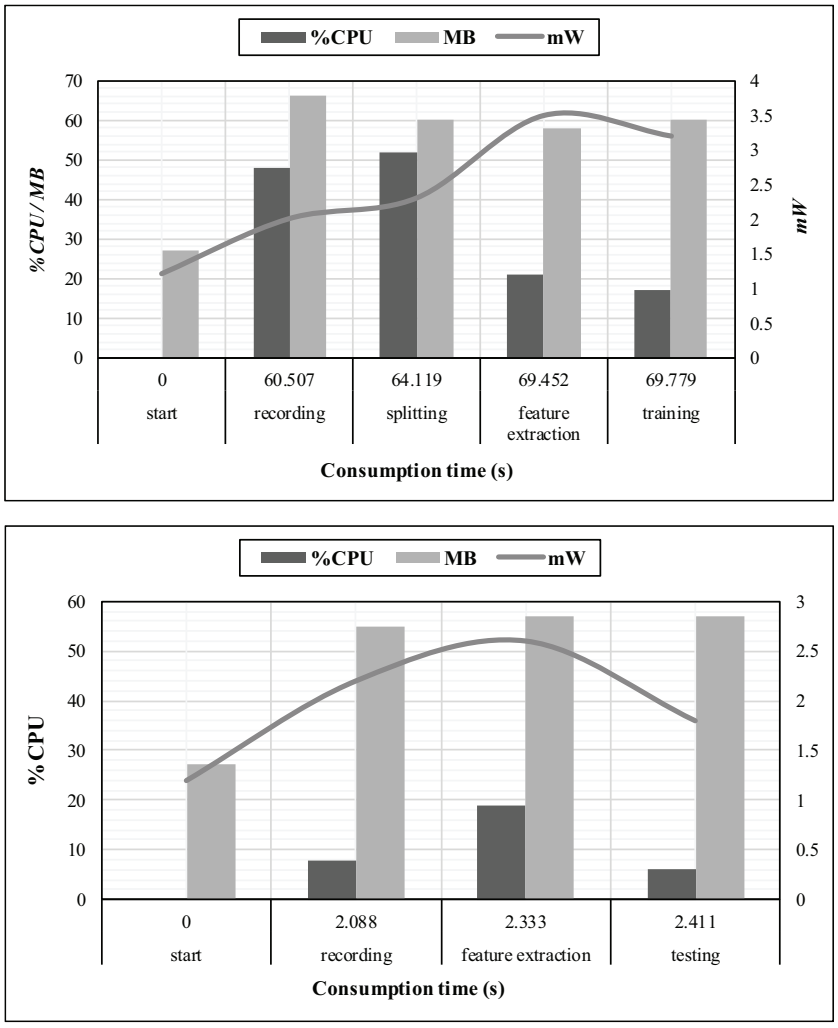

Figure 11. CPU, RAM, battery and time consumption, respectively expressed in %CPU, MB, mW and seconds, over every stage of the experiment, where the first chart refers to the training process, and the second is the identification.

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 the 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 led 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 of the nine who locked their device as well as one of 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 thinking 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 dataset is compared to the ones we have built through both the Ted-LIUM and a subset of the TIMIT corpora. Moreover, these results are 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 their DTW-based solution that involves quadratic time and space complexities. Due to the use of LPCC features, comparing our technique directly with MFCCs-based ones is very limited. However, according to the comparison detailed in Table 4, the results achieved by our proposed system also are consistent with the work of Reynolds and Rose [20]. Moreover, the reliability of our proposed system is acceptable, in real-life recording conditions, with a smaller number of classes than previous works. However, 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 participated in 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 the 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 pieces of information that mobile devices may contain. Hence, by introducing the notion of access privileges, we also aim at reducing unsafe situations in the 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 nil. Therefore, we esteem that such a technique might be a more significant option as part of a multilayer authentication. Moreover, it should also be better employed as a more reliable fall-back solution in order to eradicate PIN codes.

6. Conclusions

In this research, we have proposed the design of a TiSA system for mobile devices with a specific focus on its usability. This implementation operates as a stand-alone system that does not require any network communications. Indeed, both training and identification phases, which are based on LPCCs and the naive Bayes classifier, are achieved on 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. The results we have obtained over the different recognition analyses we have performed demonstrate the reliability and the efficiency of our TiSA mechanism in both quiet and noisy environments for a small set of persons (i.e., 90% and 80% of kappa in quiet and noisy environments,

respectively, for eleven users). In addition, the resources consumption analysis performed in this work shows the ability for the system to run on the weakest mobile devices.

We found that seven 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 sceptical participants, we also suggest that this technique should be either used in a multilayer authentication system or as a fall-back 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 LPCC features.

Acknowledgments: This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), through the discovery grant of Bob-Antoine J. Menelas Number 418624-2013. Moreover, the authors would like to acknowledge Sébastien Gaboury and Valère Plantevin for their conscientious reviews, as well as every person who participated in our experiment.

Author Contributions: The authors contributed equally.

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

Abbreviations

The following abbreviations are used in this manuscript:

ANN	Artificial Neural Network
AES	American-English Speaker
ARCSIN	Arcus Sin Coefficients
AVG	Average
CDBN	Convolutional Deep Belief Networks
CPU	Central Processing Unit
DBFS	Decibels Relative to Full Scale
DTW	Dynamic Time Warping
EER	Equal Error Rate
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LAR	Log Area Ratio
LPC	Linear Prediction Coefficient
LPCC	Linear Prediction Cepstral Coefficient
LSF	Line Spectral Frequencies
MB	Megabyte
MFCC	Mel-Frequency Cepstral Coefficient
PCM	Pulse-Code Modulation
PIN	Personal Identification Number
RC	Reflection coefficients
SPH	NIST Sphere Format
TIMIT	Texas Instruments and Massachusetts Institute of Technology

TiSA Text independent Speaker Authentication
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; Gartner Inc.: Stamford, CT, USA, 2015.
2. Wilska, T.A. Mobile phone use as part of young people's consumption styles. *J. Consum. Policy* **2003**, *26*, 441–463.
3. Goggin, G. *Cell Phone Culture: Mobile Technology in Everyday Life*; Routledge: Abingdon, UK, 2012.
4. Falaki, H.; Mahajan, R.; Kandula, S.; Lymberopoulos, D.; Govindan, R.; Estrin, D. Diversity in smartphone usage. In Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, San Francisco, CA, USA, 15–18 June 2010; ACM: New York, NY, USA, 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. In Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, Stockholm, Sweden, 30 August–2 September 2011; ACM: New York, NY, USA, 2011; pp. 465–473.
6. Yan, J.; Blackwell, A.; Anderson, R.; Grant, A. Password memorability and security: Empirical results. *IEEE Secur. Priv.* **2004**, *2*, 25–31.
7. Clarke, N.L.; Furnell, S.M. Authentication of users on mobile telephones—A survey of attitudes and practices. *Comput. Secur.* **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. *Comput. Secur.* **2002**, *21*, 220–228.
9. Yampolskiy, R.V. Analyzing user password selection behavior for reduction of password space. In Proceedings of the 2006 40th Annual IEEE International Carnahan Conferences Security Technology, Lexington, KY, USA, 16–19 October 2006; pp. 109–115.
10. Bond, R.H.; Kramer, A.; Gozzini, G. Molded Fingerprint Sensor Structure with Indicia Regions. U.S. Patent D652,332, 17 January 2012.
11. Derawi, M.O.; Nickel, C.; Bours, P.; Busch, C. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In Proceedings of the Sixth IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), Darmstadt, Germany, 15–17 October 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. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 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 Commun.* **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 and Sons: Chichester, UK, 2011.
17. Jain, A.; Bolle, R.; Pankanti, S. *Biometrics: Personal Identification in Networked Society*; Springer Science and Business Media: Dordrecht, The Netherlands, 2006; Volume 479.
18. Kumar, R.; Ranjan, R.; Singh, S.K.; Kala, R.; Shukla, A.; Tiwari, R. Multilingual speaker recognition using neural network. In Proceedings of the Frontiers of Research on Speech and Music (FRSM-2009), Gwalior, India, 15–16 December 2009; pp. 1–8.
19. Nair, R.; Salam, N. A reliable speaker verification system based on LPCC and DTW. In Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, 18–20 December 2014; pp. 1–4.

20. Reynolds, D.A.; Rose, R.C. Robust text independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 72–83.
21. Khoury, E.; Vesnicer, B.; Franco-Pedroso, J.; Violato, R.; Boulkcnafet, Z.; Fernández, L.M.; Diez, M.; Kosmala, J.; Khemiri, H.; Cipr, T.; et al. The 2013 speaker recognition evaluation in mobile environment. In Proceedings of the 2013 IEEE International Conference on Biometrics (ICB), Madrid, Spain, 4–7 June 2013; pp. 1–8.
22. Kinnunen, T.; Li, H. An overview of text independent speaker recognition: From features to supervectors. *Speech Commun.* **2010**, *52*, 12–40.
23. Thullier, F.; Bouchard, B.; Ménélas, B.A.J. Exploring Mobile Authentication Mechanisms from Personal Identification Numbers to Biometrics Including the Future Trend. In *Protecting Mobile Networks and Devices: Challenges and Solutions*; CRC Press: Boca Raton, FL, USA, 2016; p. 1.
24. Milanesi, C. *Voice Assistant Anyone? Yes Please, but Not in Public!* Creative Strategies, Inc.: San Jose, CA, USA, 2016.
25. Thullier, F. A Practical Application of a Text-Independent Speaker Authentication System on Mobile Devices. Master's Thesis, Université du Québec à Chicoutimi, Chicoutimi, QC, Canada, 2016.
26. Rabiner, L.R.; Juang, B.H. *Fundamentals of Speech Recognition*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1993.
27. Higgins, A.; Vermilyea, D. KING Speaker Verification LDC95S22. 1995. Available online: http://bit.ly/2fB8vQF (accessed on 30 August 2017).
28. LeCun, Y.A.; Bottou, L.; Orr, G.B.; Müller, K.R. Efficient backprop. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 9–48.
29. Berndt, D.J.; Clifford, J. Using dynamic time warping to find patterns in time series. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 31 July–1 August 1994; Volume 10, pp. 359–370.
30. Salvador, S.; Chan, P. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **2007**, *11*, 561–580.
31. Plantevin, V.; Menelas, B.A.J. Use of ecological gestures in soccer games running on mobile devices. *Int. J. Serious Games* **2014**, *1*, 49–60.
32. Lavoie, T.; Menelas, B.A.J. Design of a set of foot movements for a soccer game on a mobile phone. *Comput. Games J.* **2016**, *5*, 131–148.
33. Lee, H.; Pham, P.; Largman, Y.; Ng, A.Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1096–1104.
34. Reynolds, D.A. Speaker identification and verification using Gaussian mixture speaker models. *Speech Commun.* **1995**, *17*, 91–108.
35. Plieninger, A. Deep Learning Neural Networks on Mobile Platforms. Ph.D. Thesis, Technische Universitat München, Germany, 2016.
36. Movidius. Google and Movidius to Enhance Deep Learning Capabilities in Next-Gen Devices. 2016. Available online: http://bit.ly/1TirejP (accessed on 30 August 2017).
37. Rao, K.S.; Vuppala, A.K.; Chakrabarti, S.; Dutta, L. Robust speaker recognition on mobile devices. In Proceedings of the 2010 IEEE International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 18–21 July 2010; pp. 1–5.
38. Brunet, K.; Taam, K.; Cherrier, E.; Faye, N.; Rosenberger, C. Speaker Recognition for Mobile User Authentication: An Android Solution. In Proceedings of the 8ème Conférence sur la Sécurité des Architectures Réseaux et Systèmes d'Information (SAR SSI), Mont de Marsan, France, September 2013; p. 10.
39. Obuchi, Y. CMU PDA Database. 2002. Available online: http://bit.ly/2fBzQ5q (accessed on 30 August 2017).
40. Lindberg, J.; Blomberg, M. Vulnerability in speaker verification—A study of technical impostor techniques. *Eurospeech* **1999**, *99*, 1211–1214.
41. Sadjadi, S.O.; Hansen, J.H. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Process. Lett.* **2013**, *20*, 197–200.
42. Benesty, J.; Sondhi, M.M.; Huang, Y. *Springer Handbook of Speech Processing*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2007.

43. Al-Hassani, M.D.; Kadhim, A.A. Design a text-prompt speaker recognition system using LPC-derived features. In Proceedings of the 13th International Arab Conference on Information Technology ACIT, Balamand, Lebanon, 11–13 December 2012; pp. 10–13.
44. Kinnunen, T. Spectral Features for Automatic Text-Independent Speaker Recognition. Ph.D. Thesis, University of Joensuu, Kuopio, Finland, 2003.
45. John, G.H.; Langley, P. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–20 August 1995; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; pp. 338–345.
46. Zhang, H. Exploring conditions for the optimality of naive Bayes. *Int. J. Pattern Recognit. Artif. Intell.* **2015**, *19*, 183–198.
47. Rousseau, A.; Deléglise, P.; Estève, Y. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. In Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 3935–3939.
48. 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; U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology, Computer Systems Laboratory, Advanced Systems Division: Gaithersburg, MD, USA, 1993; Volume 93.
49. Ben-David, A. A lot of randomness is hiding in accuracy. *Eng. Appl. Artif. Intell.* **2007**, *20*, 875–885.
50. Villalba, J.; Lleida, E. Preventing replay attacks on speaker verification systems. In Proceedings of the IEEE International Carnahan Conference on Security Technology (ICCST), Barcelona, Spain, 18–21 October 2011; pp. 1–8.