# Speaker Recognition for Mobile User Authentication: An Android Solution

Kevin Brunet, Karim Taam, Estelle Cherrier, Ndiaga Faye, Christophe Rosenberger

# Speaker Recognition for Mobile User Authentication

K. Brunet (kevin.brunet@ecole.ensicaen.fr)*
K. Taam (karim.taam@ecole.ensicaen.fr)*
E. Cherrier (estelle.cherrier@greyc.ensicaen.fr)*
N. Faye (ndiaga.faye@ensicaen.fr)*
C. Rosenberger (christophe.rosenberger@greyc.ensicaen.fr)*

**Abstract:** This paper deals with a biometric solution for authentication on mobile devices. Among the possible biometric modalities, speaker recognition seems the most natural choice for a mobile phone. This work lies in the continuation of our previous work [1], where we evaluated a candidate algorithm in terms of performance and time processing. The proposed solution is implemented here as an Android application. Its performances are evaluated both on a public database and on an own made database. The obtained results are promising for well chosen parameters set, with an Equal Error Rate (EER) value of 4.52%.

## 1 Introduction

The tremendous development of applications on mobile phones involves more and more security needs. Particularly, authentication for mobile devices is becoming a very challenging issue. The mobile's owner can be authenticated by 1) something he/she knows, like a password or a PIN, 2) something he/she owns, like a token or a smartcard, 3) something he/she is or does, referring to a biometric modality like fingerprint, iris, face, keystroke dynamics for example. In fact, the first two methods reveal themselves to be insufficient to strongly guarantee the identity of a user, whereas biometrics provides the strongest link between the template used to login and the user. Therefore, we address in this paper a biometric solution for authentication on mobile devices.

Biometric systems involve two steps. The first step concerns the user enrolment: enrolment means first the capture of the biometric raw data, the features extraction to define a model (called reference) of each genuine user and its storage (if the template meets some quality requirements). In the second step called verification (used either for authentication or identification purpose), the user must present the same biometric modality and the new issued template is compared to the stored reference. If the difference between both templates is lower than a predefined threshold (determined and fixed by an operator), the user is authenticated and accepted by the system. Otherwise, he/she is rejected.

* 1 Normandie Univ, France ; 2 UNICAEN, GREYC, F-14032 Caen, France; 3 ENSICAEN, GREYC, F-14032 Caen, France; 4 CNRS, UMR 6072, F-14032 Caen, France

However, all biometric modalities are not suited for a smart object use: some biometric sensors are already present in the object itself, providing them with inherent biometric abilities. We can mention: a microphone, a camera, a touch screen, and for some smart devices a fingerprint reader. Therefore, the development of an authentication solution for mobile must be fitted to the already present sensors. In the sequel, we focus on the most natural biometric modality for mobile authentication, namely speaker recognition.

This paper lies in the continuation of our previous paper [1], where we addressed the problem of finding a simple solution that could be further embedded in a mobile, among the existing speaker recognition techniques. The performances of the selected method were applied to a suited database, and evaluated in terms of EER, recognition rate and verification time. We intend in this paper to show the implementation as an Android application of the previous algorithm prototype and to quantify operational performance. This paper is organized as follows. The section 2 presents the state of the art of biometrics for mobile authentication with a focus on speaker recognition. Section 3 is dedicated to a description of the proposed system, related to text-independent speaker recognition. Section 4 presents the obtained results both on an own made database and a public database using a real mobile phone. The conclusions and the perspectives are drawn in the last section.

## 2 State of the art

Biometrics refers to the identification of humans by their characteristics or traits; we can mention three categories of biometric modalities.

a) Morphological biometrics
It is based on the measure of some morphological characteristics, such as: face, fingerprint, iris, hand geometry, voice, etc...
The disadvantage of morphological biometrics is that these characteristics are time-varying and are inevitably subject to time aging.

b) Biological biometrics
It is based on the measure of some biological characteristics, such as DNA (which is the most used the modality among biological biometrics), blood, hair, etc... It is generally used in forensics applications.

c) Behavioral biometrics
It is based on users behavior. The most used behavioral modalities are: the signature, the keystroke dynamics, the gait recognition, the voice, etc...

In the literature, biometrics based mobile authentication is an emerging issue, with relatively few references. The NIST report [17] details some recommandations concerning portable biometric acquisition station and considers the following modalities: fingerprint, face and iris. In the recent paper [20], the authors propose an overview about biometrics on mobile phone through some standard modalities (fingerprint, speaker recognition, iris recognition, gait) and present a new application to ECG measurement and remote telecardiology, with an extra portable heart monitoring device. In the literature dedicated to biometric solutions for mobile authentication, most of papers are related to a specific

modality. Face recognition is dealt with in the paper [8], along with eye detection, or in [3], where a real time training algorithm is developed for mobile devices. The authors propose to extract local face features using some local random bases and then to incrementally train a neural network. Image processing also concerns hand biometrics on mobile as in the reference [6], where hand images are acquired by a mobile device without any constraint in orientation, distance to camera or illumination. The author of [11] details an iris recognition system, based on a three-step pre-processing method relying on (a) automatic segmentation for pupil region, (b) helper data extraction and pupil detection and (c) eyelids detection and feature matching. Some recent papers [4], [9], and [2] deal with keystroke based recognition. The first paper makes a study about user identification using keystroke dynamics-based authentication (KDA) on mobile devices, relying on 11-digit telephone numbers and text messages as well as 4-digit PINs to classify users. The second develops a more performant KDA process, with optimized enrolment and verification steps, whose principle is extended in the latter paper for touch screen handled mobile devices, along with a pressure feature measurement. The reference [7] presents a new modality for authentication on mobile device, namely gait recognition. The quite recent references [13] and [18] are focused on speaker verification on mobile. The first deals with text-dependent speaker verification. It means that both the user's voice and the uttered text itself are used for the verification. The second paper proposes a new method to extract features from speech spectra called *slice features*.

The aforementioned references show that many biometric modalities can be used for user authentication on mobile device. In this paper, following our previous work [1], we focus on speaker recognition. Indeed, the users are less used to take a picture of the hand, their fingerprint, or to draw a secret path on the screen of the mobile phone. Therefore, speaker recognition seems the most natural modality choice for implementation on mobile phone. In the vast literature dealing with speaker recognition, two trends stand out: text-dependent and text-independent speaker recognition techniques. We are only interested in the second field. Among the intense literature on this topic, we just refer the reader to the thorough survey paper [12] and the associated references. Using classical speaker recognition techniques to design an authentication system based on a biometric challenge on a mobile phone is not straightforward. Indeed, some constraints, inherent to the use of a mobile device, must be taken into account from the design step: the quality of the sound acquisition depends on the characteristics of the embedded microphone and the environment, the complexity of the embedded algorithms must be adapted to the capacity of the smartphone in terms of memory and processing power. The aim of this paper is to test an implementation of the algorithm presented in [1], under the form of an Android application on a mobile phone. In our previous work, we addressed the problem of finding a simple solution that could be further embedded in a mobile, among the existing speaker recognition techniques. The performances of the selected method were tested on the Sphinx Database of the Carnegie Mellon University [16], in terms of EER, recognition rate and verification time. In a second step which gives rise to the present paper, we want to compare the obtained theoretical results to performances in a real use conditions.

The next section is devoted to a description of the proposed system.

# 3 Proposed system

## 3.1 Text-independent speaker recognition

The human voice is a complex information-bearing signal, depending on physical and behavioral characteristics. The raw speech signal, uttered by any person, is extremely rich in the sense that it involves high dimensional features. To perform efficient speaker recognition, one must reduce this complexity, while keeping sufficient information in the extracted feature vector. Some speaker recognition methods have become popular in recent years: the reader is referred to the survey paper [12] for more details. Here, we briefly recall the text-independent speaker recognition process, where five steps are considered.

- Signal acquisition

Microphones and analog-digital converter are used to record and digitize the user's voice. At the end of this step, a numerical vector representing the uttered speech is available. The duration of speech recording depends on the desired accuracy.

- Speech signal preprocessing

The speech signal $x(n)$ is not a stationary signal since the vocal tract is continuously deformed and the model parameters are time-varying. But, it is generally admitted that these parameters are constant over sufficiently small time intervals. Classically, the signal is divided into frames of 25 milliseconds denoted $x_i(n)$. This division into frames leads to discontinuities in the temporal domain, and inevitably to oscillations in the frequential domain. Among the possible solutions to avoid this phenomenon, a Hamming windows is applied here.

- Feature extraction

Based on the speech signal registration and preprocessing, features are extracted to define a model corresponding to the user. Ideally, these features must be robust to intrinsic variability of the user's voice (due to stress, to disease), to noise and distorsion, to impersonation. The most widely employed methods involve short-term spectral features. We have chosen to extract MFCC (Mel-frequency cepstral coefficients) introduced by [5], which reveal to be more robust and efficient in practice.

- Speaker modeling

Once these features have been extracted on each frame, the corresponding model or template design requires a training phase. Again, we choose the VQ (vector quantization) method [19]. It is based on the LBG (LindeBuzoGray) algorithm [14]. This process allows, after clustering, to describe a voice sample by a model vector having a predefined fixed size, whatever the initial length of the signal.

- Speaker recognition

These four previous steps correspond to the user enrolment phase. In the recognition step, we consider user authentication: the system must verify if the user is the mobile's owner. For VQ based modeling (see for example [10]), the recognition test is classically performed through Euclidean distance computation between the reference template and the new captured template. Notice that the acquisition conditions could be worse in this step than in the enrolment step, where the stored model must be of high quality.

In figure 1, we present the general diagram of a speaker recognition system in enrolment (or training) and test (or verification) mode.

The figure 2 illustrates the computation of the MFCC coefficients, which is briefly

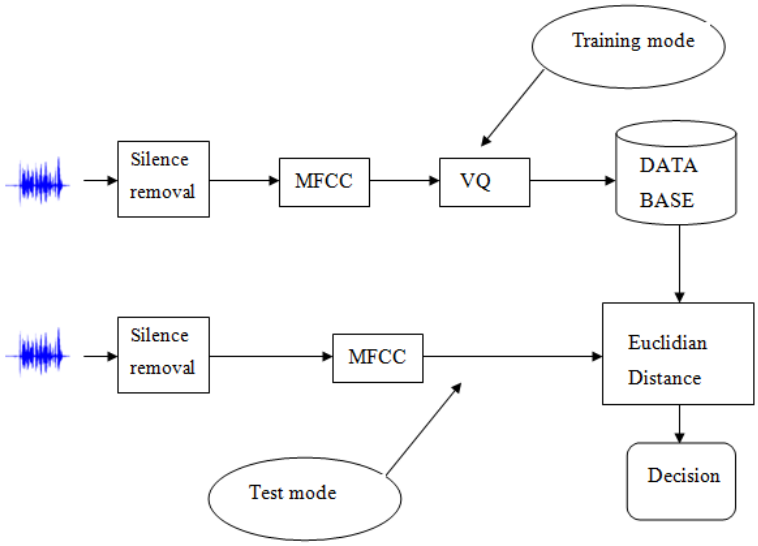



Fig. 1: General architecture of speaker recognition system

detailed below.

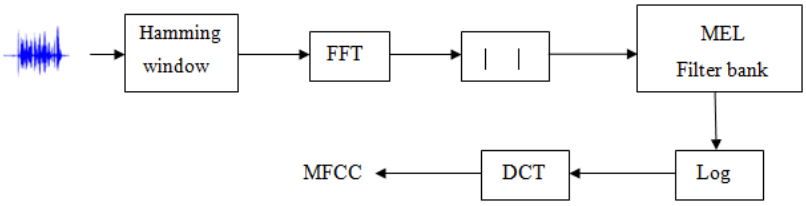



Fig. 2: Calculation process of MFCC coefficients

Consider a particular sentence denoted $x(n)$, where $n$ denotes the $n^{th}$ sample. The voice signal is divided into small frames $x_i(n)$ of 256 samples with an overlap between them of 60 %. A Hamming window is applied to each frame:

$$y_i(n) = x_i(n) * w(n) \tag{1}$$

where $y_i(n)$ is the transformed signal, $x_i(n)$ is the considered frame and $W(n)$ is the Hamming window defined by:

$$W(n) = 0.54 - 0.46.\cos(\frac{2\pi n}{256-1}) \tag{2}$$

for $0 \leq n \leq N-1$.

The Fourier transform of each frame is computed, the next step is performed in the frequency domain. The human voice spectrum is not linearly distributed, therefore, we use a Mel scale filter bank to represent the wide spectrum. A given frequency $f$ in HZ can be converted into the Mel scale [15]:

$$MEL(f) = 2595 * log_{10}(1 + \frac{f}{700}) \tag{3}$$

In general, 20 Mel filters are required for high accuracy. We apply after a logarithmic compression and a discrete cosine transform (DCT). Finally, the discrete amplitudes of the resulting cepstrum are called the MFCCs coefficients [5].

The resulting MFCC coefficients of each sentence $x(n)$ are 20 dimensional vectors. The number of vectors depend on the duration of the speech signal. A VQ process enables to reduce the dimension of data, since each vector can be represented by a given number of centroids, and then stored as reference.

In the verification step, after the extraction of the MFCC coefficients, the Euclidean distance between these parameters and the claimed reference is computed. The obtained distance is compared to a given threshold and the user is either accepted or rejected (see our previous paper [1] for more details). These different steps have been implemented as an Android application on a smartphone.

# 4 Experimental results

## *4.1 Experimental protocol*

In this section, we intend to quantify the operational performance of the proposed application. We first detail the protocol we followed. For our tests, we used two different databases:

- Sphinx database [16]: proposed by CMU (Carnegie Mellon University). It consists of voice signals collected by a PDA device. 50 sentences of about 4 to 8 seconds are uttered by 16 users. The users work at CMU, they are native speakers of American English.
- GREYC database (homemade database): it consists of voice signals collected by a Samsung Galaxy Note 1, with a version of Android that is greater than 4.0. 18 users spoke 5 times (details are given below). The data collection was realized in an office in a quiet environment.

For the GREYC database collection, the acquisition conditions are described below:

- Enrollment: the user records his/her voice on the phone for 8 to 10 seconds, then the template is generated and stored in the mobile.
- Authentication: the user speaks for 5-6 seconds and the application compares this new voice with the stored reference corresponding to the claimed user.

Now, we briefly recall how the performances of a biometric system can be evaluated. The comparison of two biometric data is a statistical process, in the sense that two captures of the same biometric data are different. The decision process (to determine if the user corresponds to the claimed identity) relies on the estimation of a similarity distance. Two classical error rates are defined as:

- the FMR (False Match Rate): it measures the probability of accepting an intruder instead of a genuine user.
- the FNMR (False Non Match Rate): it measures the probability, for a genuine user, to be falsely rejected.

These two error rates depend on a threshold, fixed by an operator, which determines the maximal distance between the stored template and the captured template from the verification step. When this threshold is varied, the error rates evolve in an opposite manner. Therefore, a particular value of this threshold, called the EER (Equal Error Rate), corresponds to the point where FMR=FNMR. In the sequel, the performances of the considered biometric system are evaluated through the computation of its EER: a good system corresponds to a low EER value.

## 4.2 Results

In a first step, an optimization of both the number of Mel-filters and the number of centroids is looked for, since the less parameters there are to define the template, the less memory is needed to store the template, and the less time is required to perform the verification. This study allows us to adjust the parameters to obtain the most efficient recognition recognition rate.

The table 1 presents a comparison of the obtained EER for the considered databases, in function of different numbers of MFCC coefficients.

| Number of Mel-filters | EER Sphinx database | EER GREYC database |
|---|---|---|
| 64 | 5.35 | 5.14 |
| 50 | 5.50 | 4.22 |
| 47 | 5.56 | 4.61 |
| 45 | 5.35 | 5.16 |
| 40 | 5.45 | 5.72 |

Tab. 1: EER for different numbers of Mel filters

It shows that it is more interesting to use 45 Mel filters for the Sphinx database and 50 when the phone's microphone is used (GREYC Database). In this second case, the obtained EER is equal to 4.22%. We choose as tradeoff to use 45 Mel filters in the rest of the study.

The table 2 presents a comparison of the obtained EER for the considered databases, in function of different numbers of centroids.

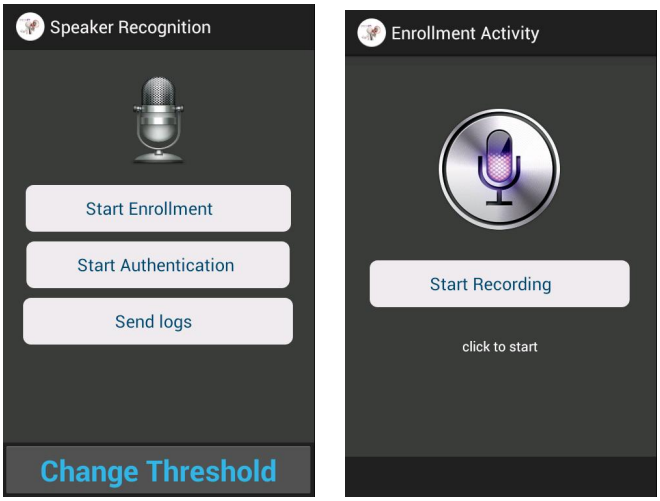
It may be noted that the EER value is lower with 256 centroids for the Sphinx database and 128 for the GREYC database. These databases do not use the same microphone, so

| Number of centroids | EER Sphinx database | EER GREYC database |
|---|---|---|
| 256 | 5.35 | 5.16 |
| 128 | 6.25 | 4.52 |

Tab. 2: EER for different numbers of centroids

the results can not be directly compared. The best obtained EER is quite promising for a first real-conditions implementation: 4.52%, with 128 centroids and 45 Mel filters.

Some screenshots of the developed Android application are shown in figure 3.




(a) Home screen (b) Enrolment screen

Fig. 3: Screenshots of the developed Android application

## 5 Conclusion and perspectives

In this paper, we addressed a biometric authentication implementation on mobile device. We choose speaker recognition modality as it seems the most natural when considering mobile phone. The proposed system is in the continuation of a previous work where we only performed theoretical results which helped us to choose a well suited algorithm, with good performances in terms of EER and time processing. The results obtained with the developed Android application, in real operating conditions are quite promising: the best EER is 4.52% for an own made database, and is better than the performances on a public database, for different choices of parameters. These results also show that the parameters and especially the number of centroids must be adapted to the considered sensor: two different mobile devices will inevitably have different parameters. Therefore, a preliminary step of calibration must be realized to get the best performances, in terms of EER and also in terms of computing time. This work is the first step in Android

application development. Further improvements will be brought, such as silence removal, noise filtering, parameter calibration.

## References

[1] M. Baloul, E. Cherrier, and C. Rosenberger. Challenge-based speaker recognition for mobile authentication. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012.

[2] T.-Y. Changa, C.-J. Tsaib, and J.-H. Lina. A graphical-based password keystroke dynamic authentication system for touch screen handled mobile devices. *The Journal of Systems and Software*, 85:1157–1165, 2012.

[3] K. Choi, K.-A. Toh, and H. Byun. Realtime training on mobile devices for face recognition applications. *Pattern Recognition*, 44:386–400, 2011.

[4] N.L. Clarke and S.M. Furnell. Advanced user authentication for mobile devices. *Computers & Security*, 26:109–119, 2007.

[5] S. Davis and P. Mermelstein. Comparison of parametric representations for mmonosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, Signal Process.*, 28:357–366, 1980.

[6] A. de Santos-Sierra, C. Sanchez-Avila, J. Guerra-Casanova, and A. Mendaza-Ormaza. *Hand Biometrics in Mobile Devices*, chapter Advanced Biometric Technologies. InTech, 2011. Available from: http://www.intechopen.com/books/advanced-biometric-technologies/hand-biometrics-in-mobile-devices1.

[7] Mohammad Omar Derawi. Biometric options for mobile phone authentication. *Biometric Technology Today*, 2011(9):5 – 7, 2011.

[8] A. Hadid, J. Y. Heikkila, O. Silven, and M. Pietikainen. Face and eye detection for person authentication in mobile phones. In *1st ACM/IEEE International Conference on Distributed Smart Cameras*, 2007.

[9] S. Hwang, S. Cho, and S. Park. Keystroke dynamics-based authentication for mobile devices. *Computers & Security*, 28:85–93, 2009.

[10] A. KABIR and S.M.M. AHSAN. Vector quantization in text dependent automatic speaker recognition using mel-frequency cepstrum coefficient. In *Proceedings of 6 th WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing, Cairo, Egypt*, pages 352–355, 2007.

[11] J.-S. Kang. Mobile iris recognition systems: An emerging biometric technology. In *International Conference on Computational Science (ICCS)*, 2010.

[12] T. Kinnunen and H. Li. An overview of text-independent speaker recogniiton: From features to supervectors. *Speech Communication*, 52:12–40, 2012.

[13] A. Kounoudes, A. Antonakoudi, V. Kekatos, and P. Peleties. Combined speech recognition and speaker verification over the fixed and mobile telephone networks. In *Proceedings of the 24th IASTED International Conference on Signal processing, Pattern Recognition, and Applications*, pages 228–233, 2006.

[14] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. Commun.*, COM-28:84–95, 1980.

[15] L. Muda, M. Begam, and I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *Arxiv preprint arXiv:1003.4083*, 2010.

[16] Y. Obuchi. Pda speech database. http://www.speech.cs.cmu.edu/databases/pda/index.html, 2002.

[17] S. Orandi and R. M. McCabe. Mobile id device. best practice recommendation. NIST Special Publication 500-280, 2009. Available from: http://www.nist.gov/itl/iad/ig/upload/MobileID-BPRS-20090825-V100.pdf.

[18] A. Roy, M. Magimai.-Doss, and S. Marcel. A fast parts-based approach to speaker verification using boosted slice classifiers. *IEEE Trans. on Information Forensics and Security*, 7:241–254, 2012.

[19] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.-H. Juang. A vector quantization approach to speaker recognition. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 1985.

[20] S. Wang and J. Liu. *Biometrics on Mobile Phone*, chapter Recent Application in Biometrics, pages 3–22. InTech, 2011. Available from: http://www.intechopen.com/books/recent-application-in-biometrics/biometrics-on-mobile-phone.