# Automatic Speaker Verification using Nearest Neighbor Normalization (3N) on an iPad Tablet

Houssemeddine Khemiri, Alexander Usoltsev, Marie-Christine Legout
Dijana Petrovska-Delacrétaz, Gérard Chollet

Institut Mines-Télécom
Télécom Sudparis-Télécom Paritech
Alexander.Usoltsev,Houssemeddine.Khemiri,Dijana.Petrovska@telecom-sudparis.eu
Mariechristine.Legout,Gerard.Chollet@telecom-paristech.fr

**Abstract:**
This paper describes the development, implementation and validation of an automatic speaker recognition system on an iPad tablet. A score normalization approach, referred as Nearest Neighbor Normalization (3N), is applied in order to improve the baseline speaker verification system. The system is evaluated on the MOBIO corpus and results show an absolute improvement of the HTER by more than 4% when the score normalization is performed. A human-centered interface is implemented for the speaker recognition system and a survey is collected from 28 users in order to evaluate the application. The results showed that the users, familiar with touchscreen interface, found the application easy to learn and use.

## 1 Introduction

Modern portable devices like tablets or mobile phones are now widely used, especially to store sensitive data (such as personal details and pictures), or to have access to email, social networks, bank accounts, etc. Most of these services require personal authentication. Therefore, biometrics could be an appropriate solution. In this paper, a biometric iPad application based on voice modality is developed. The proposed speaker verification system is based on a GMM-UBM method [RQD00], which is simpler to implement, faster and requires much less training data than an i-vector system [DKD+11]. The baseline GMM-UBM speaker verification system was evaluated during the 2013 speaker recognition evaluation in mobile environment [KVR+13]. In order to improve the performance of the baseline system, a score normalization approach based on the nearest impostor score is applied.

For the iPad application, the user interface is an important part. In fact the user interaction with the system could dramatically affect the performance of the application. Therefore an appropriate human-centered interface should be designed in order to achieve high performance level. In order to evaluate the user interface, a survey is proposed in which users are asked to assess the application interface.

The paper is organized as follows. In Section 2, the proposed score normalization approach is presented along with its evaluation on the MOBIO database. In Section 3, the implementation of the iPad application is described and the user interface evaluation results are exposed. Conclusions and perspectives are given in Section 4.

## 2 Speaker Verification System

In this section, a GMM-UBM baseline system is described and a score normalization scheme, referred as Nearest Neighbor Normalization (3N), is introduced. In the speaker verification task the goal is to decide if a person, who claims to be a target speaker, is or is not that speaker. The decision will be either an acceptance or rejection.

### 2.1 Baseline GMM-UBM System

There are two phases in any speaker verification system: the enrollment (or training) and recognition (or test). There are three main steps in these two phases. In the feature extraction step, which should be the same in the enrollment and recognition phases, the speech signal is converted into a sequence of vectors. Each vector represents a short window of the waveform with overlapping adjacent windows. Then, the speaker modeling step creates a model for each of the speakers' voices using samples of their speech. Finally, the speaker model once trained allows to perform the recognition step by scoring the test speech against the model of the claimed speaker. This score is compared to a threshold to decide whether to accept or reject the claimant.

The well-known GMM-UBM speaker verification system is proposed as a baseline system. The UBM (Universal Background Model) is a GMM with 512 Gaussians trained on a large number of speakers to represent the distribution of the extracted features for a representative population. The target speaker model is obtained by adapting the parameters of the UBM model using the target training speech with Maximum A Posteriori (MAP) adaptation. In the decision making step, a log-likelihood ratio for a test sequence of feature vectors X is computed as follows:

$$LLR(X, target) = log(p(X|\lambda_{target})) - log(p(X|\lambda_{UBM})) \quad (1)$$

Where $\lambda_{target}$ and $\lambda_{UBM}$ are respectively the target and UBM models.

### 2.2 3N Normalization Method

The speaker verification problem received a lot of attention, leading to many public evaluation campaigns such as NIST (http://www.nist.gov/itl/iad/mig/sre.cfm), MOBIO [KVR+13], Biosecure [PDCD09], etc. In this paper, a score normalization method is performed to improve the performance of the GMM-UBM speaker verification system. For the baseline GMM-UBM speaker verification system, feature extraction is first performed on the test utterance. Then a Log-Likelihood Ratio (LLR) is computed, as shown in equation 1, between the claimed identity and the test utterance. Finally the LLR value is compared with a threshold. If the LLR is above the threshold, the test speaker is accepted otherwise he is rejected.

For the 3N Normalization Scheme, in addition to the comparison with the claimed identity model, the test utterance is compared with other speaker models present in the database. Then, the difference between the LLR of the claimed identity and the maximal LLR found for the other speakers is computed as follows:

$$D(X, target) = LLR(X, target) - \max_{1 \leq i \leq N-1} (LLR(X, Impostor_i)) \quad (2)$$

If the difference is above a threshold, the test speaker is accepted otherwise he is rejected.

### 2.3 Experimental Setup

In order to validate the proposed normalization approach, the publicly available MOBIO database is used. Then the speaker verification system is evaluated on a smaller database acquired with the iPad (these results are reported in Section 3.4).

The MOBIO database is a bimodal (face/speaker) database recorded from 152 people. The database has a gender ratio of nearly 1:2 (100 males and 52 females). More details on this database could be found in [KVR+13]. Based on the gender of the clients, two different closed set protocols for male and female are proposed. In order to have an unbiased evaluation, the clients are split up into three disjoint sets: training, development and evaluation sets:

- **Training set**: The data of this set is used to learn the UBM parameters.

- **Development set**: The data of this set is used to tune meta-parameters of the algorithm (e.g. number of Gaussians, number of features, etc.). For the enrollment of a client model, 5 audio files of the client are provided. The remaining audio files of the clients serve as probe files, and likelihood scores have to be computed between all probe files and all client models. Number of trials is 60480 for male and 34020 for female.

- **Evaluation set**: The data of this set is used for computing the final evaluation performance. It has a structure similar to the development set. Number of trials is 151620 for male and 42000 for female.

In order to set the influence of the numbers of impostors of the 3N normalization on the speaker verification results, a new evaluation protocol, called full MOBIO protocol, is created by adding the MOBIO development set to the evaluation one. In such case, the number of 3N normalization speakers is increased from $N = 20-1$ to $N = 38-1$ for female, and $N = 38-1$ to $N = 62-1$ for male.

The framework for reproducible results of the baseline speaker verification system is described in [PDCD09] and available at [HPDF+09]. This framework was adapted to the MOBIO database as follows. First, the feature vector is composed of 20 MFCC coefficients (32 Mel filter bank) together with their first derivatives and the delta energy. This is intended to better exploit the 16KHz range. Second, feature warping and energy-based voice activity detection are performed. The UBM model is trained on the MOBIO training part and the Voxforge dataset (http://www.repository.voxforge1.org/) with 512 Gaussians. Moreover, the 3N score normalization approach is used. Open source SPro 4.1 [Gra09] and ALIZE 2.0 [BSM+08] softwares were used to develop the proposed system.

### 2.4 Results with the Proposed 3N Normalization Scheme

Figure 1 and 2 show the DET curves and the Half Total Error Rate (HTER) obtained for baseline system and 3N normalization, respectively for the female and male **evaluation set of the MOBIO database**. To compute the HTER, a threshold $\theta$ is defined on the development partition at the EER point. This threshold is applied to the evaluation partition to obtain the HTER. The results clearly reveal the contribution of the 3N normalization method to improve the performance of the baseline speaker verification system: 3N HTER is 7.01% for female and 4.87% for male. In fact, the absolute improvement made in the HTER is nearly 4% for female and male. It is also important to note that the results achieved by the 3N normalization are better than those obtained by the best system in the 2013 speaker recognition evaluation in mobile environment [KVR+13], which are 10.67% for female and 7.07% for male. Furthermore, the proposed system performs as good as the fusion on the all systems participating in that evaluation campaign, where the best results is 6.73% for female and 4.63% for male [KGSM13].


Figure 1: DET curves obtained for the baseline system (dotted line) and 3N normalization (continuous line) for the **female** MOBIO **evaluation** dataset (left) and **full** MOBIO dataset (right), and the HTER operating point (circle).


Figure 2: DET curves obtained for the baseline system (dotted line) and 3N normalization (continuous line) for the **male** MOBIO **evaluation** dataset (left) and **full** MOBIO dataset (right), and the HTER operating point (circle).

In order to test the influence of the size of the 3N normalization population, the speaker verification system is evaluated on the full MOBIO protocol. For the female trials, an HTER of

10.52% is obtained for the baseline system on the full MOBIO scenario. Increasing the size of the normalization population from 19 to 37 improves the speaker verification results from 7.01 to 6.23%. For the male trials, an HTER of 8.98% is obtained for the baseline system on the full MOBIO scenario. Increasing the size of the normalization population from 37 to 61 slightly deteriorates the speaker verification results from 4.87 to 5.01%. Therefore, more tests should be done to verify influence of the size of the normalisation population.

## 3 Implementation, Architecture and User Interface

In Section 3.1, the importance of a human friendly interface for biometrics systems is discussed. Then in Sections 3.2 and 3.3, the architecture and the interface of our biometric application are explained. Finally results of our user interaction experiments are provided in Section 3.4.

### 3.1 Human Interface for Biometric Systems

One of the problems in building biometric systems is to establish a good contact with the end user. As reported by [oST08] in the beginning stages of development of biometric systems, it was necessary to focus primarily on the performance. As these new technologies mature, it's important to take into account other factors, including the usability of these systems.

One aspect of biometric systems that has not been traditionally considered is the user. The user brings innate qualities and experiences to the interaction that affect performance. Without proper human-centered interface, biometric systems can't receive biometric modality of good quality, decreasing the performance of the biometric system. According to ISO 9241 [92410] human-centric design is an iterative process and should involve: early focus on users, tasks, and environment; active involvement of users; appropriate allocation of tasks between user and system; incorporation of user-derived feedback into the (biometric) system design; iterative design whereby a prototype is designed, tested and modified.

### 3.2 Architecture and Implementation

In the case of voice biometrics the user pronounces a sentence and a biometric system process that speech to perform the biometric comparison. To solve the problem of human-centered interface, the popular mobile platform - iOS - is chosen to implement our biometric system on an iPad device. Using iPad as an end-user terminal seems quite natural because of its touchscreen interface, big screen and sensitive microphone. Also, current iPad Air model line has powerful hardware [hiar] to run both enrolment and test phases right on the device in a reasonable time. To develop an application which will fit users needs, it is important to decide what this application will do. The main function of our application is a login function. Whenever registered user starts application and wants to have access to some protected data in iPad, he or she provides a sample of his/her biometric modality. The application grants or denies access to protected data. For example, it could be a doctor who wants to have an access to patients medical data.

For our biometric application on iPad, a classic Model-View-Controller software design pattern is chosen. It is a high-level pattern that deals with the global architecture of an application and classifies objects according to the general roles they play in an application [hVCVC].

Logical application architecture consists of three parts: iPad hardware (camera, microphone); Objective-C classes to perform the speaker verification (speech acquisition, feature extraction and processing, speaker verification); Objective-C classes to store data (file-based database to keep all users' feature templates as a binary files).

### 3.3 Application Interface

For the proposed biometric application on the iPad, a very first prototype of the human interface is developed with separated processes for enrollment and test. While both of the processes follow similar path "record speech - extract features - process features", from the interface point of view they are different. Normally, user needs enrolment phase only once and usually he is supported by the application guide. The test process, where user claims his identity and provide his biometric modality, is most frequently used.

A few initial tests with real users have shown that the human interface acceptance could be improved by cutting out unnecessary steps. Apple's "iOS Human Interface Guidelines" [Dev14] recommend to avoid interface parts which do not bring any functionality for the user, and therefore make it easy to focus on the main task by highlighting important content or functionality. Since the user will spent most of his time using the "test phase", one extra interface element is removed and the application starts immediately with the test phase interface. When needed, the user can switch to the enrollment interface. With that simpler version of the interface, more user tests are performed and a survey is proposed in order to evaluate the interface.

### 3.4 User Interface Survey

All test related to the proposed user interface are conducted on a population of 28 users to assess the usability of the biometric application. Usability testing focuses on scenarios related to enrollment and verification and aims to address the difficulties encountered when using the application. The proposed survey is inspired by the criteria proposed in [SB97] to evaluate the results. The questionnaire aims to gather the subjective assessment of the application for the user. The obtained results are reported in Table 1. During testing, the number of mistakes, the completion time for assessing the efficiency and the effective achievement of objectives are reported. After the test, a questionnaire is given to gather subjective assessment of the device, difficulties, misunderstandings, obstacles and what they like or not in the system.

A small amount of data are acquired with the iPad from the 28 users, leading to 419 genuine and 3307 impostors tests. This data is considered as a development corpus to fix the decision threshold on the Equal Error Rate (ERR) oprating point. The baseline system achieves an EER of 5.23%, while by adding the 3N approach to normalize the scores, the EER decreases to 1.22%. More data will be collected form more users and tests. This data will be exploited to evaluate the performance of the implemented application by using the threshold previously computed on the development part.

| | Questions | 2 | 1 | 0 |
|---|---|---|---|---|
| Global impression | Global impression of the application | 21 | 6 | 1 |
| | Impression of the application: easy-to-use | 25 | 1 | 2 |
| | Impression of the application: flexible | 18 | 7 | 3 |
| | Impression of the application: interesting | 23 | 3 | 2 |
| | Impression of the application: satisfaction | 24 | 3 | 1 |
| Screen | Easy to read text on a screen | 25 | 1 | 2 |
| | Information clearly organized | 22 | 3 | 3 |
| | Position and color of the a buttons helps to complete task | 20 | 8 | 0 |
| | Screen sequence is clear | 19 | 9 | 0 |
| Terminology | Instructions are clear | 18 | 5 | 5 |
| | Messages well positioned on a screen | 21 | 4 | 3 |
| | Appropriate use of terms in the application | 24 | 2 | 2 |
| Ergonomics | Application is easy to learn | 26 | 2 | 0 |
| | Always easy to complete task | 25 | 3 | 0 |
| | Messages on the screen are useful | 25 | 3 | 0 |
| | Steps and instructions are easy to remember | 23 | 5 | 0 |
| Capabilities | Application is strict | 18 | 9 | 1 |
| | Application is running fast | 27 | 1 | 0 |
| | Can be used by any user | 17 | 10 | 1 |
| | Errors are easy to correct | 15 | 5 | 6 |

Table 1: Survey Results on 28 users ( 2 = totally agree, 1 = agree and 0 = not agree)

## 4 Conclusions and Perspectives

The proposed speaker verification application demonstrates significant progress in two main directions. First, it showed that 3N normalization brings an absolute improvement of the HTER of the speaker recognition system to more than 4%. An other important result is that modern portable devices (like iPad tablet) are mature and powerful enough to run speaker recognition algorithms based on GMM-UBM method in real-time. Second, human interface surveys were performed on a group of users and surveys results showed that the users, familiar with tactile interface or iPad interface, found the application easy to learn and use. That user reaction is one of the important factors to build a high-performance biometric system.

Future works will be dedicated to implement face recognition on our application and to improve the performance of the system by fusing scores between speaker and face recognition. Furthermore, several tests will be conducted in order to calibrate the decision threshold. In addition, good scores on the criteria of speed and learning obtained in the surveys should be validated by testing with bigger group of users in operational environment that reflect the contexts of use.

## Acknowledgements

This work is partially supported by the FUI 15 Equip'Age project.

## References


[92410] ISO 9241. Ergonomics of human-system interaction, provides guidance on human-system interaction throughout the life cycle of interactive systems, 2010.

[BSM+08] J.F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason. ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. In *The Speaker and Language Recognition Workshop, Odyssey*, 2008.

[Dev14] Apple Developper. iOS Human Interface Guidelines, 2014.

[DKD+11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011.

[Gra09] G. Gravier. SPro: Speech Signal Processing Toolkit, release 4.1, 2009.

[hiar] http://www.anandtech.com/show/7460/apple-ipad-air review/3.

[HPDF+09] A. El Hannani, D. Petrovska-Delacrétaz, B. Fauve, A. Mayoue, J. Mason, J.F. Bonastre, and G. Chollet. Text-independent Speaker Verification. In *Guide to Biometric Reference Systems and Performance Evaluation*. Springer, 2009.

[hVCVC] https://developer.apple.com/library/ios/documentation/general/conceptual /CocoaEncyclopedia/Model View-Controller/Model-View-Controller.html.

[KGSM13] E. Khoury, M. Günther, L. El Shafey, and S. Marcel. On the Improvements of Uni-modal and Bi-modal Fusions of Speaker and Face Recognition for Mobile Biometrics. In *Biometric Technologies in Forensic Science*, 2013.

[KVR+13] E. Khoury, B. Vesnicer, J. Franco-Pedroso R., Violato, Z. Boulkcnafet, L.M. Mazaira Fernandez, M. Diez, J. Kosmala, H. Khemiri, T. Cipr, R. Saeidi, M. Gunther, J. Zganec-Gros, R.Z. Candil, F. Simoes, M. Bengherabi, A. Alvarez Marquina, M. Penagarikano, A. Abad, M. Boulayemen, P. Schwarz P., D. Van Leeuwen, J. Gonzalez-Dominguez, M.U. Neto, E. Boutellaa, P. Gomez Vilda, A. Varona, D. Petrovska-Delacretaz, P. Matejka, J. Gonzalez-Rodriguez, T. Pereira, F. Harizi, L.J. Rodriguez-Fuentes, L. El Shafey, M. Angeloni, G. Bordel, G. Chollet, and S. Marcel. The 2013 speaker recognition evaluation in mobile environment. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, June 2013.

[oST08] National Institute of Standards and Technology. Usability and Biometrics, 2008.

[PDCD09] Dijana Petrovska-Delacrétaz, Gérard Chollet, and Bernadette Dorizzi. *Guide to Biometric Reference Systems and Performance Evaluation*. Springer Verlag, 2009.

[RQD00] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(13):19 – 41, 2000.

[SB97] D.L. Scapin and J.M.C. Bastien. Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour and Information Technology*, 16(4-5):220–231, 1997.