Journal of Engineering Science and Technology Review
Special Issue on Telecommunications, Informatics, Energy and Management 2019

JOURNAL OF
Engineering Science and Technology Review

Conference Article

www.jestr.org

Multifactor Voice-Based Authentication System

Marian Ceaparu[1], Stefan-Adrian Toma [3], Svetlana Segarceanu[1], George Suciu[1] and Inge Gavat[2]

[1]Beia Consult International, R&D Department, Bucharest, Romania
[2]Politehnica University of Bucharest, Department of applied Electronics and information engineering Bucharest, Romania
[3]Military Technical Academy, Bucharest, Romania

Received 29 September 2019; Accepted 20 February 2020

Abstract

The paper addresses an important nowadays technological tendency: interaction with customers through voice interfaces in order to improve the services towards them. It proposes a framework for a user authentication application on smart phones, as a preamble for any intelligent system operating speech technology. Authentication is based on the mobile device and a number of factors derived from speech, actually involving a voice based biometric solution. The article will describe the important components of such a system, and early results of the research. One important step is the design of an operating scenario for system interaction with customers and the conversation outline at enrolment and exploitation. The authentication system involves a Speaker Verification (SV), an Automatic Speech Recognition (ASR) component, and a decision fusion logic: their combination makes what is known as prompt-based verification. The output of the dialog at enrolment, consisting in user utterances of her/his identification information, is used for training the SV system, by generating individual speakers' models and contributing to the background model. At exploitation the customer, identified by the characteristics of her/his mobile device, will have to answer some questions proposed by the system, based on the speech material at enrolment, and/or utter some digit sequences. These utterances are processed by the SV system and the ASR. The SV outputs a first authentication factor. The ASR identifies the text uttered by the unknown user, and thus provides another number of factors. An overall fusion rule will be introduced to merge all these factors.

Keywords: Prompt Speaker Verification: Speech Recognition: Fusion Rule: Multiple Factors Authentication> Voice based Interfaces.

1. Introduction

An important technological trend in information and communication systems is the employment of voice-based interfaces to facilitate interaction with customers and improve the services towards them. This tendency is evident as hundreds of millions of devices using voice as interface are available today as shown in Opus Research Reports [1].

Any intelligent system operating speech technology requires a voice-based authentication, a biometric solution intended to substitute traditional physical authentication means such as passport, ID card (something you have)) for access control. The benefits of biometrics (something you are) are conspicuous in remote transacting, which traditionally relied exclusively on personal information or passwords (something you know). However, the performance of biometric solutions is limited by several issues, including environmental factors, like noise for voice, compromise by attackers, security inconveniences such ast he impossibility of reconfiguring [2]. As any single-factor security system, biometric systems can be inefficient given these shortcomings. Although they are generally more difficult to attack than systems based on passwords, they can still be vulnerable. Therefore, many authentication systems consider several factors, among which biometrics.

The biometric factor derived from voice is called voice print and the process involved in voice biometrics is Speaker Verification (SV). The goal of an SV system is to either confirm or to invalidate the pretended identity of an unknown speaker based on her/his speech characteristics. From the point of view of the text uttered by the user, SV applications can be classified as: text dependent, ·text independent and prompt based. Text dependent SV operates with fixed passwords, established usually by users. Text independent SV systems use any text utterances. Prompt based SV relies on text items proposed by the system. Some of these approaches perform besides anSV task, an automatic speech recognition (ASR) task. Voice biometrics can be active when the user states knowingly the required information. It can be passive, in the background of a conversation with the agent, of the unaware customer.

In the last 20 years the SV discipline had a fast evolution. The first systems to offer a useful degree of accuracy were the GMM – UBM (Gaussian Mixture Modelling – Universal Background Model) systems [3] [4], based on simple statistical models for users and universal background models of speakers. Since this approach was developed, Machine Learning (ML) progressed rapidly, mainly thanks to the availability of huge audio data

*E-mail address: svetlana.segarceanu@beia.ro
ISSN: 1791-2377

resources.

Since then many ML approaches have been applied to the problem, either in isolation or in combination, including: Support Vector Machines (SVMs) [5], Eigenvector Analysis [6], Joint Factor Analysis (JFA) [7], Linear Discriminant Analysis (LDA), Probabilistic Linear Discriminant Analysis (PLDA)[8] [9]. Until recently, the state-of-the-art approach was I-Vectors [10], a combination of some of these techniques. Latterly they have been overtaken by X-Vectors [11], which also draw on the techniques listed above.

However, there is another major drawback concerning voice biometrics: it is not as accurate as other biometric modalities, for instance as fingerprint or facial recognition. For this reason, most present-day solutions develop multi-modal user authentication. Among these compound solutions worth mentioning Nuance biometricauthentication products merging voice and face recognition12], AWARE [13] that provides the framework for several biometric solutions and fusion instruments for different modalities.

The present paper introduces a voice-based authentication framework to be implemented as an application on a smart phone. It can stand as a preamble to any intelligent system operating speech technology and requiring voice-based authentication. The proposed authentication approach is based on multiple factors derived from speech: the voiceprint and the uttered text. A third factor would be the very device of the customer.

One concern is the ease-of-use and friendliness of this system, a second issue is its performance. Therefore, on the one hand, we introduce a set of scenarios to be applied at enrolment and authentication. On the other hand, we describe the methodologies involved in setting up this system. The decisions of SV and ASR are merged according to a fusion rule. Speakers' training will be performed using the speech acquired at enrolment. Speech recognition is trained to recognize sequences of digits (based on an existing speech material) and certain keywords which will be learned progressively. The scenario concerns the description of the front-end involving the speech material, and interaction based on the dialogue between the user and the system. The application will be tailored to be applied in Romania.

2. User Interface for a biometric authentication system

The interface design involves two distinct stages: enrolment and authentication.

2.1 Enrolment

At enrolment the client will be asked to provide through her/his device regular identification information available on the identification card: name, address, ID number, date of birth, Social Security Number (SSN). Moreover, the user might be asked to provide additional control information such as alternative telephone number, occupation, names of the parents. In Romania a unique identification code, called CNP (numerical personal code) is used. It represents a 13-digit long sequence, encrypting information about the owner: gender, date of birth, residence, a security encoding.

2.2 Authentication

At authentication the client will be identified by the characteristics of her/his device. Such a characteristic might be the phone number of the owner. The application on the device will prompt on the screen some random text and/or a number of questions to verify the identity of the speaker claimed by her/his device. The questions will be created at random based on the speech material provided by the user at enrolment; the random text would be a sequence of five or six digits. We propose a set of three questions: the first one might beuser's name or phone number. The other combination of two prompted questions could be:

1. The year of birth and the last six digits of the CNP
2. The name of the father, and date of birth
3. The number of the identification document and the name of the street where she/he lives.
4. Street name and a random sequence of digits.

The answers to these questions will represent inputs to equally the SV and the ASR modules.

3. Speech-based Authentication System

The operating diagram of the system in the authentication step is presented in figure 1. The main component of the system is the Speaker authentication module. This module receives as input two or more speech recordings of the user and the claimed identity of the speaker invoked through the mobile device and decoded by a phone identification service.

The speaker authentication module includes two components receiving speech as input:

1. Speaker Verification
2. Speech Recognition

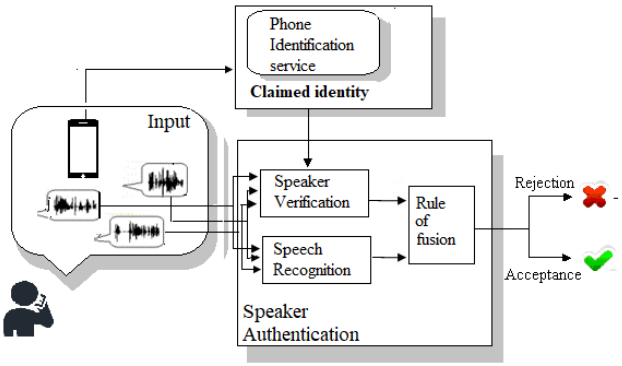

Fig 1. Speaker Authentication Operating Diagram

SV and ASR are based on three input speech waves. The decisions of the two modules will be merged into a conclusive decision, by a fusion engine. We describe henceforward the components of the system.

3.1 Speaker Verification

The task of SV is to decide on the invoked identity of a user: either client or impostor. It is a 1:1 evaluation.

We have implemented an SV system usingt he sheer GMM –UBM approach, with maximum a posteriori (MAP) adaptation [3]. The reason for choosing the GMM-UBM is that, as noticed in [14], under certain conditions of limited training data and short utterances, a standard GMM-UBM SV system may achieve better performance than an i-vector/PLDA based system.

SV process, typically involves the following levels:

1. Training
2. Testing and calibration
3. Adaptation
4. Verification

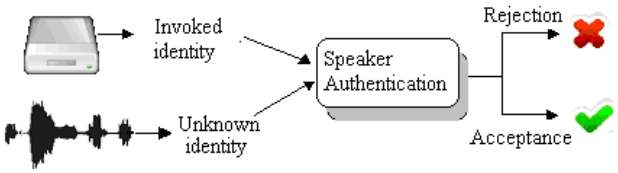

Fig 2. Speaker Verification diagram

3.1.1 Training

In the GMM-UBM approach,t raining means generating:

- GMM models for each individual speaker
- GMM model for the UBM

that is generating the parameters of the entities to be modelled, given a number K of components: mean-μ_k, standard deviation -Σ_k, weight-c_k for model component *k,* using the Expectation-Maximization (EM) algorithm on the enrolment data. GMM models $\lambda_i = \{(c_{ik}, \mu_{ik}, \Sigma_{ik}\}, 1 \leq k \leq K\}$ for each speaker *i*, use12-28 components. Modelling is applied to the extracted features space; we tested three types of features: Mel-frequency Cepstral Coefficients (MFCC) [16], Linear Prediction Cepstral Coefficients (LPCC) [17] and derived from Perceptual Linear Prediction (PLP) [18].

The UBM model λ_{ubm} is generated from all the acquired material from speakers participating in the experiment and other available speech resources. The recommended number of components for the UBM [3] is something between 1024 and 2048. The $\lambda_{ubm}$is generated as sum of several GMM sub-models, each one with K_k components, for Msub populations of speakers [4] [15], so that:$\lambda_{ubmk} = \{(c_{ubmki}, \mu_{ubmki}, \Sigma_{ubmki}), 1 \leq i \leq K_k, 1 \leq k \leq M\}$.

3.1.2 Verification

The SV process, as shown in Figure1, is based on three factors, the utterances of the required prompt texts of the client user, as shown in section 2.Each verification process is text-independent so that each utterance is evaluated against the same speaker model and background model. The evaluation of each of these utterances against the claimed identity of speaker *i* is made by the test of the ratio between the likelihood rates, or difference of the log-likelihood rates:

$$\Lambda(X) = \log P(X/\lambda_i) - \log P(X/\lambda_{ubm}) > \theta_i \qquad (1)$$

where θ_i is the specific threshold of user *i*, but in fact, it is often a common value for all the users. *X* stands for the extracted feature vector.

To integrate the evaluations of the three utterances we average the scores of the utterances X_k and compare the mean to the speaker's threshold, as outlined in the formula:

$$\Lambda(X_1, X_2, X_3) = \frac{(\sum_{k=1}^{3}(\log P(X/\lambda_i) - \log P(X/\lambda_{ubm})))}{3} > \theta_i \qquad (2)$$

For an overall evaluation of anSV system performance, two error rates are estimated, given a threshold value.

- **FAR** (False Acceptance Rate) which accounts for all the impostor speakers that are accepted by the system.
- **FRR** (False Rejection Rate) -which accounts for all the user speakers that are incorrectly rejected by the system.

Usually the threshold value for which the two error rates are roughly equal (Equal Error Rate-**EER**) is used. Yet, other operating points for the threshold can be set, depending on how important is one or the other of the two rates.

3.1.3 Calibration and Adaptation

This section discusses two issues. The first problem concerns setting the individual thresholds. For each user the scores obtained on a test set of impostor utterances, claiming the respective speaker identity, are considered. An approach is to apply score normalization [18]for the formula in (2), using score mean and standard deviation.

$$\bar{\Lambda}(X) = \frac{\Lambda(X) - \mu_\lambda}{\sigma_\lambda} > \theta_i \qquad (3)$$

For normal distribution of impostor scores, setting the threshold value θ_i to 1.5insures theoretically a FAR of100-86.638 = 13.362%(13.362% of impostors are likely to be accepted).

The common value 1.5 of the threshold could be used unless a proper sensitivity term s_i is added for each user. The individual values for sensitivity or threshold can be set by estimating the error rates for a range of threshold values and pick the value for the suitable error rates ratio. This approach will be referred to as *norm1*. The best-known normalization approach is *znorm*, where the θ_i estimation in (3) is set to 1.

Another important issue is constant model adaptation, by using newly acquired speech from already authenticated users, on the one hand to fill up the models of the user speakers. On the other hand, we apply MAP adaptation to adjust (some of)the GMM parameters of UBM.

The experimental section will present some early experimental results for SV [20] using one or two factors, in text-semi-independent, or text independent SV employing the three types of characteristic features mentioned above.

3.2 Speech Recognition

Currently the ASR component is implemented using the open source solution Kaldi [21].This toolkit was chosen for its good identification rate and speed. It provides several alternative approaches to ASR and we have tested two solutions: the first one using MFCC as characteristic features and Hidden Markov Models (HMM)[22] as modelling method, the second one using the MFCC-Deep Neural Networks (DNN) [23]. This component was trained on two corpuses of Romanian speech to provide:

- recognition of sequences of 10 digits, and a range of isolated words in Romanian
- recognition of a range of words from a bank specific conversation in continuous speech

Several indicators are used to evaluate an ASR:

- **Identification Rate** estimated as number of correctly identified words against the total number of uttered words;
- **Word error rate-WER** calculated as number of erroneous words that appear in the transcription of the vocal signal, against the total number of uttered words. An

erroneous word might be either an omitted, or a substituted word or an inserted word;

- **Confusion matrices** are useful to highlight the confusions between words, or different linguistic units. Kaldi accounts for all types of errors met at automatic speech recognition, Confusion errors are deduced from all the substitutions registered by Kaldi.

Because the application requires a more complex ASR solution, for Romanian, we propose using a Dynamic Time warping (DTW) solution, for the specific limited vocabulary or a hybrid solution as proposed in [24] [25].

3.3 Speech Material

Speech related tasks were tested using specific corpuses. The ASR function was tested on several speech corpuses. The first corpus, CORPUS1contains the speech of 29 male and 12 female speakers, who uttered sequences of digits and some seven words in Romanian, amounting to 130 audio files for each speaker. The corpus was recorded in an anechoic room, the audio files were sampled at 16 kHz. Because a speech recognition system is useful in decoding telephone communication, these recordings are not suited for a valuable evaluation. To add a channel effect to the audio files, they were retransmitted through the telephone network, by an iterative process, resulting in a second corpus CORPUS2. It is assumed that CORPUS1 is included in CORPUS2. Three types of telephonic connexions were envisaged, linking mobile phones and land line. To simulate real environment conditions natural background noise was added. For ASR testing and evaluation tasks the two corpuses were split into corresponding subsets, for training and testing, in a proportion of 3:1. ASR tests relate to two models created based either on CORPUS1 or CORPUS2. The tests using CORPUS1 for modelling and testing are labelled TEST1, tests using CORPUS1 for modelling and andCORPUS2 for testing are labelled TEST2, and tests relying on CORPUS2 for training and testing are called TEST3. In all these tests modelling and classification are accomplished using the monophone MFCC-HMM paradigm. Other tests were performed on CORPUS2, using the trip hone MFCC-HMM, and DNN modelling.

Another important collection of speech material, in Romanian, was intended to mimic a bank call-centre application. It contains more than 39 hours of recordings at a sampling frequency of 16 kHz. About10% of the collected speech contains predefine text. Speech records hold corresponding transcriptions and all the relevant information to generate the linguistic and acoustic models. Evaluation of the ASR modules was accomplished using the testing functionality of Kaldi.

For the SV task, we used a corpus containing speech of 14 female and 12 male speakers, who pronounced a set of six compulsory sentences and arbitrary text in Romanian, throughout 4 to 11 recording sessions. 21 of them were used as client speakers. For text-independent SV this resulted in 1757 authentic speaker and 45624 impostor utterances.

At present we are collecting speech material to comply with the requirements of the proposed system in what concerns the uttered text, and well balanced from the point of view gender, age, etc.

3.4 Fusion Rule

The decision upon the identity of the unknown speaker is based on the decisions of the SV module and the ASR system. Both applications rely upon three utterances of the unknown speaker. The SV module itself uses a specific fusion rule, explained by relation (2), to combine the decisions based on different utterances, and furnish a merged decision *decision1*.

As shown in figure 3 *decision1* should be merged with the outputs of the ASR module produced by the three utterances. We propose the following fusion rule:

$$\text{final decision} = \text{decision1} \wedge (\text{decision2} \vee \text{decision3} \vee \text{decision4}) \quad (4)$$

This rule states that the biometric authentication is compulsory, even though based on merging of several SV decisions, and at least one of the three uttered texts should be correctly identified.

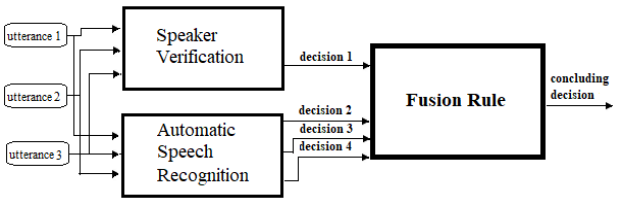

Fig 3. Fusion of several authentication factors

3.5 Phone Identification Service

Calling will be initiated though an application explicitly designed for this purpose. Along with user enrolment, device enrolment is performed, i.e. registering the device with the server. Registration consists of generating a public private key pair, linked to the phone IMEI number. Prior to calling, a signed request is sent to the server. After receiving a validation message, the phone can initiate the call.

4. Experimental Results

This section presents the results for speaker verification (4.1) and ASR evaluation (4.2).

4.1 Speaker Verification

We have performed three types of SV tests, using either (one or two) passwords, set identically for all speakers, or speaker dependent (tables 1 and 2), or arbitrary text (Table 3). The experimental results show that using two passwords can soar the EER performance by some 4 percent. Tests using several random texts were not performed, but it is expected that the EER decreases by some percent as compared to using one only text. We used the basic GMM-UBM, MAP adaptation of all parameters (UBM_adapt) or of some parameters (UBM_adapt_wm), score normalization, or removal of not relevant UBM components. This part was developed in Java.

4.2 ASR Evaluation

This section presents the results of five types of tests on CORPUS1 and CORPUS2 and the confusion matrices for two of these tests, in figures 4 and 5.

Table 4 presents the Identification rates and the WER for the five types of tests presented in section 3.3. Obviously, the results obtained using the ideal recordings produced the best performance of 99.7%, and a very low word error rate, even using the monophone paradigm, applying MFCC-HMM approach. Evaluating the ASR on the corpus processed through telephone network produced the best

performances when applying the trip hone paradigm and the MFCC-HMM approach. The performance of the DNN ASR is very little lower.

Table 1. Verification EERs for experiments using one password, either *password1* or *password2* common to all speakers, using several features, score normalization and UBM adaptation

Password	Test Name	LPC	MFCC	PLP
Password1	Norm1	8.70	13.90	15.20
	Znorm	10.00	14.50	15.20
	UBM_adapt – Norm1	**10.35**	**11.40**	**9.75**
	Basic	12.20	14.00	11.45
Password2	Norm1	8.25	13.10	13.50
	Znorm	9.75	14.36	13.30
	UBM_adapt – Norm1	**9.20**	**9.25**	**11.69**
	Basic	10.10	10.50	12.90

Table 2. Verification EERs for two password scheme, (one shared password *password1* or *password2*, a second user specific) using several characteristic features, score normalization and UBM adaptation

Password	Test Name	LPC	MFCC	PLP
Password1	Norm1	**4.55**	7.15	**9.75**
	Znorm	5.15	7.15	9.90
	UBM_adapt – Norm1	7.00	7.90	11.50
	Basic	7.05	**7.05**	11.40
Password2	Norm1	**4.00**	8.30	**9.90**
	Znorm	3.99	7.96	10.30
	UBM_adapt – Norm1	8.05	7.87	9.95
	Basic	8.60	**7.60**	10.03

Table 3. Text Independent SV EERs using several characteristic features, score normalization, and the UBM adaptation method

Test Name	LPC	MFCC	PLP
Basic	13.40	14.20	15.45
Norm1	13.50	14.90	15.45
Znorm	12.20	14.80	15.76
UBM_adapt – Norm1	13.63	14.33	15.51
UBM adapt – Znorm	13.00	14.65	15.50
UBM_adapt_wm– Norm1	13.60	**13.44**	**15.10**
UBM_adapt_wm– Znorm	12.60	14.70	15.60
UBM-removal-Norm1	12.65	14.28	15.50
UBM-removal-Znorm	**11.90**	14.50	15.16

The confusion matrices in figures 4 and 5 stress the high confusion between the words "șase" and "șapte" (six and seven) and a certain word ("Urgență" - Emergency), taken for many other words.

On the other hand, experiments performed with DTW produced very good results on a small vocabulary, and satisfactory results for a larger speech corpus

Table 4. Synthesis of ASR evaluation for all types of tests described in Section 3.4, on CORPUS1 and CORPUS2. It shows that the ideal conditions provide obviously the best performance. Using the same type of recordings at training and testing improves the identification rate. Applying the trip hone model also soars the performance. Applying the state-of-the-art technique DNN has not produced better results.

Test Name	Identification rate	WER
TEST1	99.7	0.03
TEST2	71.68	38.47
TEST3	90.36	12.57
Triphone HMM	95.84	7.08
DNN	94.90	7.46

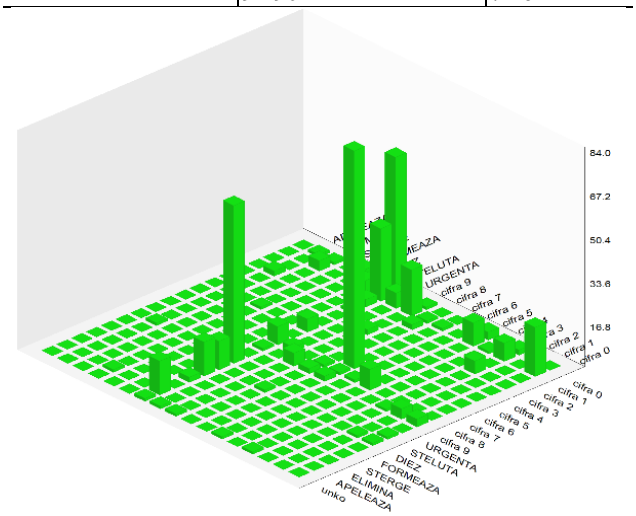

Fig 4.Confusion matrix obtained by applying the triphone MFCC-HMM modelling. The confusion of "șase" (six) taken as "șapte" (seven) is conspicuous. Moreover,the word "Urgență" is confounded with many other words.

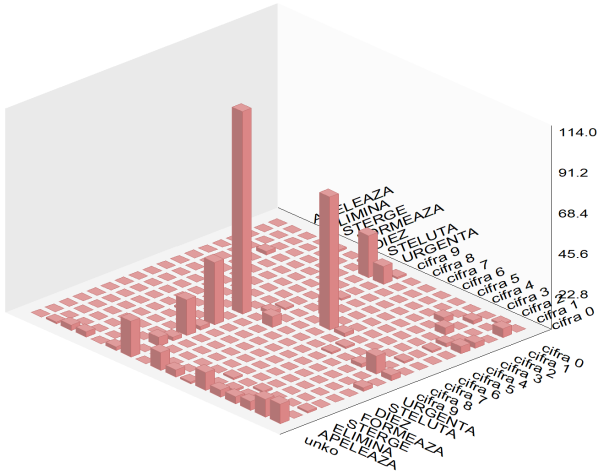

Fig 5.Confusion matrix obtained by applying the triphone DNN modelling. The confusion of "șase" (six) taken as "șapte" (seven) is also outstanding, and "Urgență" is confounded especially with "zece" (ten).

5. Conclusions

We have proposed a conceptual scheme for a voice-based authentication system. The proposed system relies on a biometric component (Speaker Verification), an Automatic Speech Recognition component, a fusion rule for heterogeneous factors and a mobile device identification module. Some of the functionalities were implemented and tested using specific tools and speech material. However, they should be integrated into one application and soundly tested on relevant speech material.

The aggregate of SV, ASR and fusion engine in figure 1, represents in fact the prompt-based speaker verification scheme. Speaker verification applies the text-independent scheme.

For the future we intend to implement, test and evaluate some state-of-the-art methodologies for the text-independent SV task, using for instance deep embedding of speakers. Another exigency concerns the implementation of a "softer"

fusion rule, applying a weighted decision logic, or maybe a probabilistic logic, as shown in [26].

Acknowledgements

The paper „Multifactor voice-based authentication system" has been funded in a great measure by EUREKA-EUROSTARS Programme: E!9870 - Speech2Process _S2P (Smart, Natural Language Semantic Analyzer Platform to Process Oriented Back-ends). The work has been supported in part by UEFISCDI Romania under grants Speech2Process, Nr.49E/2015.

Many thanks to our colleagues in BEIA Consult International, Elena Olteanu and Lucian Necula and all others to have supported our work with their advices, and voices.

References

1. https://opusresearch.net/wordpress/2019/05/06/dont-draw-the-wrong-conclusions-about-voice-biometrics-and-hmrc/.
2. https://www.sestek.com/2016/11/advantages-disadvantages-biometric-authentication/.
3. Reynolds D., Quatieri T., and Dunn R., "Speaker Verification using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, no. 1-3, pp. 19–41, 2000
4. Bimbot, F. et. al., "Tutorial on Text-independent Speaker Verification ", EURASIP Journal on Applied Signal Processing 2004:4, 430-451
5. "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation" W. M. Campbell, D. E. Sturim, D. A. Reynolds, A. Solomonoff, Proceedings of ICASSP, 2006
6. P. Nguyen, R. Kuhn, J.-C. Junqua, N._Niedzielski .C. Wellekens "Eigenvoices: A compact representation of speakers in model space", annals of telecommunications - annales des télécommunications 55(3):163-171
7. P. Kenny, "Joint factor analysis of speaker and session variability: Theory and Algorithms, Tech Report CRIM-06/08-13," 2005, Online: http://www.crim.ca/perso/patrick.kenny
8. P. Kenny, "Bayesian speaker verification with heavy tailed priors," Proc. Odyssey Speaker and Language Recogntion Workshop, Brno, Czech Republic, June 2010
9. N. Brummer, "EM for Probabilistic LDA," Available online at: https://sites.google.com/site/nikobrummer, Feb. 2010.
10. D. Garcia-Romero, X. Zhou, Y. Espy-Wilson, "Multicondition Training of Gaussian PLDA Models in i-vector Space for Noise and Reverberation Robust Speaker Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4257-4260, 2012.
11. D. Snyder, D. Garcia-Romero, G. Sell,, S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," Conference: ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
12. https://www.nuance.com/omni-channel-customer-engagement/security/identification-and-verification.html
13. https://www.aware.com/voice-authentication/
14. F. Curelaru "Evaluation of the Standard i-Vectors-based Speaker Verification Systems on Limited Data", COMM201, Bucharest, June 2018
15. Rosenberg, A. E. and Parthasarathy, S., "Speaker background models for connected digit password speaker verification", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, May 1996, pp. 81–84.
16. Davis, S. B. and Mermelstein, P Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on ASSP 28, 357–366. . (1980)
17. Markel, JD, AH Gray Jr, Linear Prediction of Speech, Spring Verlag - New York, 1976
18. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. Jour. of ASA 87(4), 1738–1752. (1990).
19. Matsui, T. and Furui, S., "Likelihood normalization for speaker verification using a phonemeand speaker-independent model", Speech Commun. 17 (1995), 109–116.
20. Svetlana Segărceanu, Tiberius Zaharia, Text-Independent Speaker Verification Using the GMM-UBM Modelling, SpeD2013, Bucharest October 2013.
21. http://kaldi-asr.org/doc/about.html
22. Lawrence R. Rabiner "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE 77.2, pp. 257-286, 1989
23. https://www.ibm.com/developerworks/library/cc-machine-learning-deep-learning-architectures/index.html
24. Tiberius .Zaharia, Svetlana Segarceanu, Marius Cotescu, Inge Gavat, Alexandru Spataru, "Binary reduction methods for reference templatesused in DTW algorithm for speech recognition in mobile applications", SISOM 2010 and Symposium of the Commission of Acoustics, Bucharest, May 2010
25. Svetlana Segărceanu, Tiberius Zaharia, "Speaker verification using the dynamic time warping", UPBScientific Bulletin, January 2013
26. Richardson, M., Domingos, P. "Markov logic networks" Mach. n. 62 (2006), 107–136.