Voice Biometrics Based on Pitch Replication

L.C. Moreno and P.B. Lopes
Universidade Presbiteriana Mackenzie,
Programa de Pós-Graduação em Engenharia Elétrica e Computação
São Paulo, São Paulo, Brasil
lsmoreno@uol.com.br ; 71717595@mackenzie,com.br

Abstract:

Authentication and security in automated systems have become very much necessary in our days and many techniques have been proposed towards this end. One of these alternatives is biometrics in which human body characteristics are used to authenticate the system user. The objective of this article is to present a method of text independent speaker identification through the replication of pitch characteristics. Pitch is an important speech feature and is used in a variety of applications, including voice biometrics. The proposed method of speaker identification is based on short segments of speech, namely, three seconds for training and three seconds for the speaker determination. From these segments pitch characteristics are extracted and are used in the proposed method of replication for identification of the speaker.

Keywords: authentication, biometrics, pitch, algorithm, pitch replication

1-Introduction

Two challenges affect speaker recognition algorithms: the diversity of channels used to transmit the speech signal, in training phase or recording test (example: phone call, cell phone, TV channel, internet, etc.) and additive noise contamination. Several techniques have been proposed to minimize or even eliminate these adverse effects. In addition to these two main problems mentioned, the duration of training and test time has a considerable effect on the performance of the speaker's recognition [1-2-3].

Research in speaker identification systems has used training and test time values of approximately 5 minutes. However in practice, this procedure becomes cumbersome because it requires the manipulation of a large amount of data, either for training or actual execution. There are several studies that analyze the performance of speech recognition systems for different recording times [4-5-6]. Within this concept, analysis were performed to determine the shortest time for recognizing a speaker, and the minimum value of 10 seconds was suggested [8-9-10],

The Figure 1 show the performance of SV accuracy in EER (%)[26] for test with segments of different lengths. The training duration is fixed at~2.5 min [10].

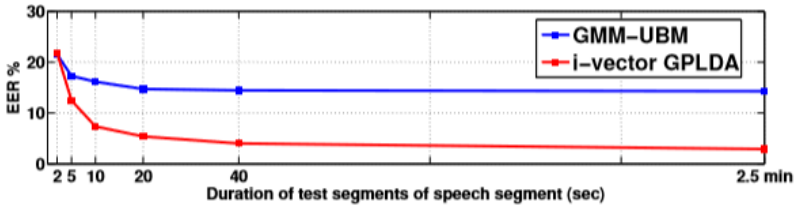

Figure1 – Performance of speaker, different length [10]

The method here proposed is to use three seconds of recording for each speaker to be identified, and to obtain recognition rates compatible with biometric access control systems in controlled environments. The proposal is to maximize the number of pitch estimates using three extractors for each speech recording, and replicate them for defined number of previously defined times to form a training base for later comparison and classification by means of the k nearest neighbor (k-NN) method. The proposed replication model will be tested in an automatic speaker identification system (Speaker Identification) using a previously recorded voice database.

2. A Brief Review of Speaker Recognition Systems

Automatic speaker recognition systems (ASR) have essentially three stages [11-12]: the first step is the transformation of the characteristics of the analog signals into digital signals, [13], the second step is to model and extract specific parameters that uniquely represent the characteristics of the given speech recording [14] and the last step consists in parameter classification to identify the speaker.

For a long time, the speaker identification used GMM (Gaussian Mixture Models) as a classifier to train the characteristics of the voice in low intensity vectors [15-16]. The Gaussian and Universal background model (GMM - UBM) are another popular approaches for text independent speaker recognition, because it allows management of both speaker and an impostor patterns, UBM has the characteristic to deal with features "hidden" [17]. The state-of-the art uses *i-vector* models and techniques of JFA (Join Factor Analysis) for speaker identification [7-19-25].

As for the extraction of voice features, the proposed model is based on Pitch. The fundamental frequency (pitch) varies from person to person and depends exclusively on the size of the vocal cord, its flexibility, its physiological quality and how it is physically structured in the larynx. The pitch value for an adult man can range from 50 Hz to 250 Hz and for women and children can reach values of 500 Hz [22]. Speaker can control the pitch of the sound being produced because the system is all supported by muscles and cartilage that can be altered through muscular contractions. The difference in fundamental frequency values between different speakers and groups of speakers has been seen as a great potential for automatic speech recognition. Pitch is considered a weak feature for a method of speaker recognition, compared to more advanced techniques such as cepstral features, Mel-Frequency cespstral coefficients (MFCC) and PMCC (Power-Normalized Cepstral Coefficients) techniques that may reflect vectors of order N, but very susceptible to noise [20]. The proposed method is the speaker identification using pitch replication comparisons and classifier k-NN (k nearest neighbors) [21].

The biometry process basically contemplates two variables that change constantly, one of them is the variation of the extracted features and the other is the means or method used to obtain the information, called the channel. In the case of voice biometry, the quick voice variations, whether in the amplitude, cadence, pronunciation, physical or emotional conditions of the speaker, etc., change the values of the features extracted for comparison, and the another one is the variation of the media for capturing the signal, such as noisy environments, telephone line, transmission medium equipment among others, which directly affect the information are called channel characteristics [22]. These two aspects, represent challenges for voice biometry and identification of the speaker as the classification deals with N-dimensional parameter vectors, located in a hyper-plan that are distinct but morphologically grouped[18]. Techniques such as i-

vector, GMM-UBM and deep neural networks, known as DNN (deep neural networks) are considered as state-of-the-art in speaker recognition systems.

3. Proposed method

The speaker recognition system proposed in this project aims to classify patterns of an speaker, using the k-NN (k nearest neighbor) technique. The methodology uses parameters to represent a specific characteristic of speech, in this case, pitch values, obtained in the stage of preprocessing of the voice signal, for the generation of temporal matrices. These temporal matrices reproduce the global and local variations in time, as well as the spectrum of the signal. Pitch replication is used to increase the amount of information that is supplied to the k-NN classifier. The performance of the proposed search algorithm is analyzed. The recording and processing of voice for speaker identification is done through a platform developed in MATLAB.

a) The proposed algorithm is best explained by the four actions described below:Initially, , as mentioned, the speech signal was partitioned into three seconds segments.
b) Then the values and pitch quantities of the three second segments are obtained using three proposed extractors algorithms: Cepstrum method (CEPSTRAL) [20-22], Subharmonic-to-Harmonic ratio (SHR) [24] and Maximum value of Fast Fourier Transform (MFFT) [22].. Due to the characteristics of each extractor, the amount of useful pitch obtained by extractor differs in extracted quantities and pitch values. On average, each extractor can obtain between 30 and 80 useful pitches values (in a recording of 3 seconds) that will be used in the proposed model of replication and later classification. It was adopted as useful pitch values, those that oscillate between 30% of the average value of the pitch samples of a recording of three seconds. As observed, due to the variations of the characteristics extracted from a speaker, as well as the variation of the communication channel, the quantity and values of pitch change, even if the same utterance was recorded.
c) Thirdly, the proposed replication technique is applied. The method consists of replicating the values and quantities of pitch values obtained from the second stage until reaching a predetermined maximum amount of pitch values. For example, if the SHR extractor obtained 50 pitch values from a given 3-second recording and the maximum number of replicate pitch values to be analyzed is 600 pitch (for example), the SHR 50 values will be replicated 12 times, thereby creating a matrix 1x600 (50 x 12 = 600).When set, the maximum pitch quantity (is maintained for all recordings (test and training recordings) and for all proposed extractors, in the case (CEPSTRAL, SHR and MFFT). There will be times when the maximum amount is not an integer number, so the value will be either truncated or completed until the maximum value proposed is reached. Following the previous example, if the MFFT gets 70 useful pitch values of a given recording and the maximum value for comparison is 600 pitch, we will have the following replication model: 8 x 70 = 560 pitch values and will be completed with 40 more pitch values of the 70 obtained in the MFFT extractor until reaching the maximum value of 600 pitch values.
d) The fourth, and final step, consists in identifying the speaker by determining the vector in the trained vector set that is closest to the actual speech data that is to be identified, and the proximity between the trained database recordings and the test database recordings, using a k- NN (k nearest neighbors) for each pitch . This will define the speaker candidate with the highest probability of being the one that recorded the speech file under analysis.

Figure 2 presents in a simplified way the method proposed of pitch replication obtained from the three extractors (CEPSTRAL, SHR and MFFT), for comparison and classification using k-NN.

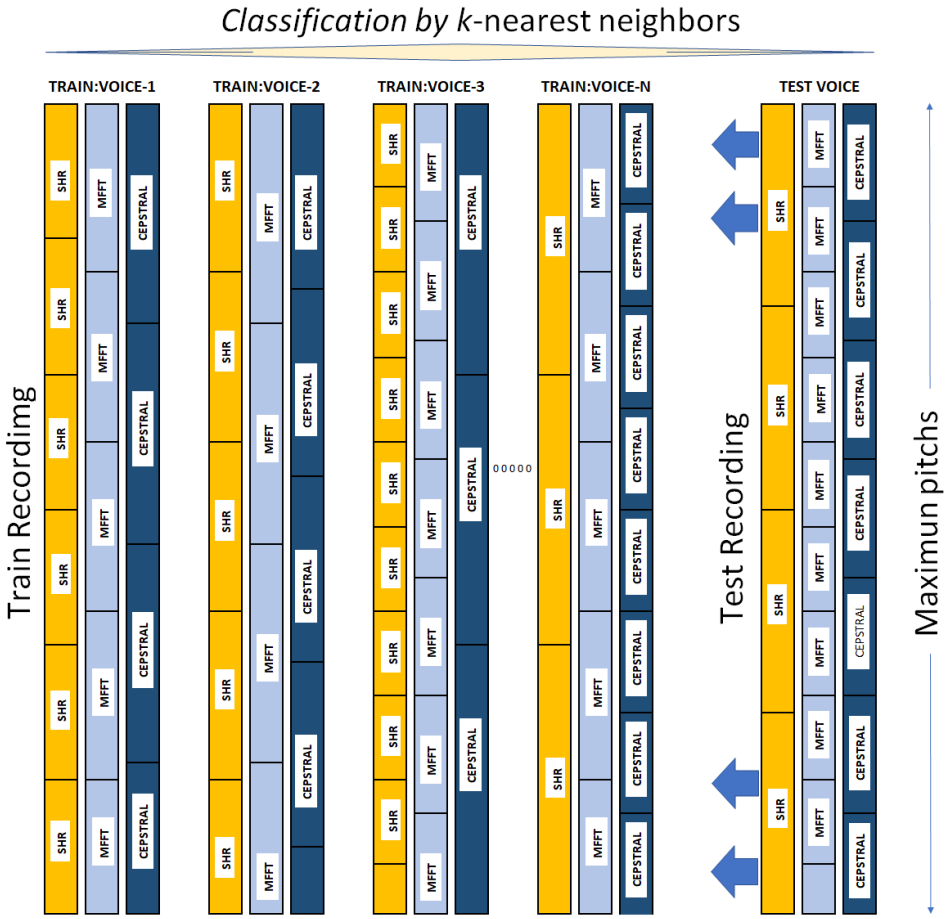

Figure2 – Comparative method of pitch replication and classification

4. Experimental Results

The proposed model of replication was tested as automatic speaker identification systems (Speaker Identification -SI) using the training and test recordings of the ELSDSR database to verify the replication method, were used 154 training recordings and the 22 test recording.

The English Language Speech Database for Speaker Recognition (ELSDSR) was prepared in the University of Denmark. The texts are in English language and are read by 20 Danes, one Icelandic and one Canadian. All the users make seven recordings, of seven different utterances, totaling 154 recordings, These recordings are considered as a training base. For the test base, the same users read two different texts, totaling 22 recordings for the test phase [23]. All recordings have an average of 6 seconds of recording, as the proposal is to use three seconds, every recording loaded for analysis was truncated in three seconds.

The experiment was divided into four phases:

1- Using the training voice database as test database, 154 recordings (7 recordings x 22 users) were used to train the algorithm and the same 154 recordings were used to test the system, using the proposed of replication for speaker identification.

2- Using part of the training voice database and the remaining use as test database. 88 (4 recordings x 22 users) of the 154 recordings were used for training purposes and the 66 (3 recordings x 22 users) remaining recordings were used for testing the algorithm
3- Using the 154 recordings (7 recordings x 22 users) of the database for training purposes and using the 22 recordings of the test database to assess the performance of the algorithm(all recording with three seconds of duration).
4- Identical to number 3 but continuously. Several recordings of three seconds continuously, until the end of the file recording of the user in the test database. In average each recording has 25 seconds of duration.

For the first experiment, due to the use of a k-NN classifier its result was 100% accuracy, all recordings were found and there wasn't FAR (False Acceptance Rate) or FRR. (False Acceptance Rate)[26].

For the second and third experiment, two analyses were performed, with replication of 600 pitchs and another with 200 pitchs. There was an improvement of 19,96% just because it performed more replication, as shown in Figure 3 (a) and 3 (b). The EER% values obtained are in accordance with speaker verification with short utterances[1-10].

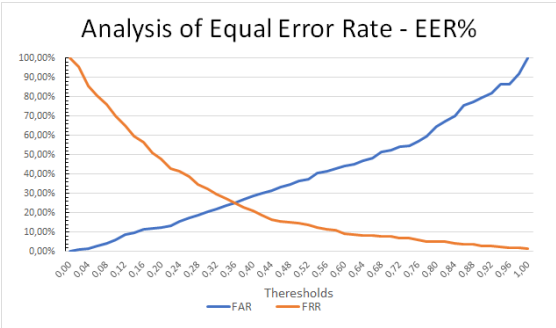

Fig.3(a) EER- 25,38% for 200 Replication

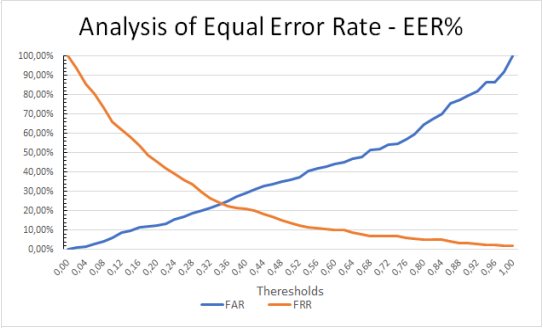

Fig.3(b) EER-22,35% for 600 Replication

For the fourth experiment, it was used 600 pitch replications. The analysis was made with all recording, that is, several samples of three seconds of the same recording passed by algorithm until the end of the recording file and an performance analysis being done at the end. Of the 22 recordings in the test database, 16 speakers were recognized and 6 speakers were not (although they were pointed out as candidates, they were not unequivocally identified as the speaker).

5. Conclusion

Speaker based identification or authentication system is becoming more popular day by day. Noise, extractor, features and channel or media differences are the basic challenges of speaker identification.

The voice biometry process basically contemplates two variables that change constantly: extracted features and channel differences. Both are directly linked to the classification models and their performance needs to be mathematically sound, computationally fast and accurate

For the proposed method, the number of pitch values to be used is augmented by a replication technique. It is demonstrated that increasing the number of replicated values of pitch tends to improve the performance

of the speaker recognition algorithm when a k-NN classifier is employed. Neural network, and Deep Learning can also be used as classifier but k-NN provides a faster training time along with a satisfying accuracy [27]. One of the main challenges in the project was to stabilize the feature values, so that there are no comparisons of segments in which the utterance is not voiced, that is a pich is not valid by definition.

Although having some challenges, voice based speaker recognition has been getting widely acceptance day by day due to some proprieties which are absent in other biometric features. Here, voice features are stored and matched without the use of noise elimination techniques and using a small amount of training data.

Experimental results confirm the accuracy of the algorithm here proposed.

6. References

[1] CemalHanilçi and FigenErtaş, "Investigation of the effect of data duration and speaker gender on text-independent speaker recognition", Computer and Electrical Engineering-2012. http://cs.uef.fi/sipu/pub/kanervisto17_gender.pdf

[2] Mak MW, Hsiao R, Mak B. A comparison of various adaptation methods for speaker verification with limited enrollment data. In: Proc. ICASSP; 2006. p. 929-32 http://www.eie.polyu.edu.hk/~mwmak/papers/icassp06Poster.pdf

[3] Vogt R, Sridharan S. "Experiments in session variability modelling for speaker verification". In: Proc. ICASSP; 2006. p. 897-900 https://dl.acm.org/citation.cfm?id=2464271

[4] Fauve BGB, Evans NWD, Pearson N, Bonastre JF, Mason JSD. "Influence of task duration in text-independent speaker verification". In: Proc. interspeech; 2007. p. 794-7. https://dl.acm.org/citation.cfm?id=2464271

[5] Vogt R, Baker B, Sridharan S. "Factor analysis subspace estimation for speaker verification with short utterances". In: Proc. interspeech; 2008. p.853-6. https://www.researchgate.net/profile/Figen_Ertas/publication/235995473_Investigation_of_the_effect_of_data_duration_and_speaker_gender_on_text-independent_speaker_recognition/links/5b090f0caca2725783e63547/Investigation-of-the-effect-of-data-duration-and-speaker-gender-on-text-independent-speaker-recognition.pdf

[6] Vogt R, Lustri C, Sridharan S. "Factor analysis modelling for speaker verification with short utterances". In: Proc. speaker Odyssey; 2008 https://eprints.qut.edu.au/12629/

[7] Vogt R, Pelecanos JW, Scheffer N, Kajarekar SS, Sridharan S. "Within-session variability modelling for factor analysis speaker verification". In: Proc. interspeech; 2009. p. 1563-6. https://www.isca-speech.org/archive/interspeech_2009/i09_1563.html

[8] Pelecanos J, Chaudhari U, Ramaswamy G. "Compensation of utterance length for speaker verification". In: Proc. speaker Odyssey; 2004. https://repositorio.uam.es/bitstream/handle/10486/7508/42232_gonzalez_dominguez_javier.pdf?sequence=1

[9] McLaren M, Vogt R, Baker B, Sridharan S. "Experiments in svm-based speaker verification using short utterances". In: Proc. speaker Odyssey; 2010. p. 83-90.

https://www.researchgate.net/publication/324936163_Improving_the_performance_of_GPLDA_speaker_verification_using_unsupervised_inter-dataset_variability_compensation_approaches

[10] ArnabPoddar, MdSahidullah, GoutamSaha. "Speaker verification with short utterances: a review of challenges, trends and opportunities". In: The Institution of Engineering and Techology 2015. https://www.researchgate.net/profile/Arnab_Poddar/publication/320201024_Speaker_Verification_with_Short_Utterances_A_Review_of_Challenges_Trends_and_Opportunities/links/5a0e84f4aca27244d2859732/Speaker-Verification-with-Short-Utterances-A-Review-of-Challenges-Trends-and-Opportunities.pdf

[11] Kinnum,T,Li,H: "An overview of tex-independent speaker recognition from features to supervectores". Speech Commun, 2010 52(1), pp12-40 http://www.cs.joensuu.fi/pages/tkinnu/webpage/pdf/speaker_recognition_overview.pdf

[12] Campbell, J.P.Jr: "Speaker recognition a tutorial", Proc. IEEE, 1997, 85(9) pp 1437-1462 https://www.lsv.uni-saarland.de/fileadmin/publications/non_articles/Speaker_Recognition_A_Tutorial.pdf

[13] John G. Proakis (Autor), Dimitris G. Manolakis: "Digital Signal Processing", 4th edition, 2007 https://engineering.purdue.edu/~ee538/DSP_Text_3rdEdition.pdf

[14] Tomi Kinnunen, Haizhou: An Overview of Tex-Independent Speaker Recognition: from Features to Supervectors", 2011 – https://hal.archives-ouvertes.fr/hal-00587602

[15] D.A. Reynolds, T.F.Quatieri, and R.B. Dunn.: "Speaker verification using adapted gaussian misture models". Digital signal processing , vol.10 no.1-3 pp 19-41.2000 https://scholar.google.com.br/scholar?q=Speaker+verification+using+adapted+gaussian+mixture+models&hl=pt-BR&as_sdt=0&as_vis=1&oi=scholart

[16] D.A. Reynolds and R.C. Rose.: "Robust text-independent speaker identification using gaussian mixture speaker models". IEEE transactions on speech and audio processing, vol.3 no1 pp-7283, 1995. https://scholar.google.com.br/scholar?q=Robust+text-independent+speaker+identification+using+gaussian+mixture+speaker+models&hl=pt-BR&as_sdt=0&as_vis=1&oi=scholart

[17] Yi-Hsiang Chao, Wei-Ho Tsai and Hsin-Min Wang: "Improving GMM-UBM speaker verification using discriminate feedback adaptation" Computer Speech&Language, 2009 https://www.researchgate.net/publication/224364035_Discriminative_Feedback_Adaptation_for_GMM-UBM_Speaker_Verification

[18] S.S. Tirumala, R.Wang.: "Speaker Identification Features Extration Methods: A Systematic Review", An International Journal on Expert Systems with Applications vol.102.July 15,2017 http://www.massey.ac.nz/~rwang/publications/17-ESwA-Reza.pdf

[19] Join Factor Analysis and i-vector Tutorial, disponivel no http:/www1.icsi.Berkeley.edu/Speech 27.03.2018. http://www1.icsi.berkeley.edu/Speech/presentations/AFRL_ICSI_visit2_JFA_tutorial_icsitalk.pdf

[20] M. T. S. Al Kaltakchi and W. L. Woo and S. S. Dlay and J. A. Chamber: Study in Fusion Strategies and Exploiting the Combination of MFCC and PMCC features for Robust Biometric Speaker Identification, 2016. https://pdfs.semanticscholar.org/86f7/488e6ad64ad3a7aab65f936c9686aee91a1a.pdf

[21] LeandroA Silva,S. M Peres Sarajane,BoscarioliClodis: Introdução a Mineração de Dados, Brasil, 2016.

https://www.loja.elsevier.com.br/introducao-a-mineracao-de-dados-9788535284461.html
[22]Lawrence Rabiner and Biing-Hwang Juang: Fundamentals of Speech Recognition, EUA 1993. https://www.amazon.com/Fundamentals-Speech-Recognition-Lawrence-Rabiner/dp/0130151572
[23]L. Feng: Speaker Recognition Informatics and Mathematical Modelling – Technical Univeristy of Demmark, DTU, ResearchGate, Sep.2004. https://www.researchgate.net/publication/259333765_Speaker_Recognition
[24]X. Sun,:A pitch determination algorithm based on subharmonic-to-harmonic ratio, pp.679-679 -6th Internacional Conference of Spoken Language Processing -China – 2000 https://ieeexplore.ieee.org/document/5743722
[25]Nayana,P,Mathewa,D and Thomasa,A:Comparation of Text Independent Speaker Identification Systems using GMM and i-Vector Methods, 2017 https://www.sciencedirect.com/science/article/pii/S1877050917318823
[26]Ruud M. Bolle,Jonathan H. Connel, Sharath Pankanti,Nalini K. Ratha and Andrew W. Senior: Guide to Biometrics, 2004 https://www.springer.com/gp/book/9780387400891
[27]Garrett Thomas, CS PhD student at Stanford: How does KNN classification compare to classification by neural networks?,2017 https://www.quora.com/How-does-KNN-classification-compare-to-classification-by-neural-networks