Journal of the Nigerian Association of Mathematical Physics
Volume26, (*March, 2014*), pp 455 – 460

Effect of Variation in Speech Characteristics on Speaker Recognition System

Emagbetere J. O.

Dept. of Electrical/Electronic Engineering,
University of Benin, Benin City

Abstract

This paper presents speaker recognition with varied speech characteristics using the Euclidean distance as an efficient classifier for speaker authentication. The speaker recognition model used is a text-dependent speaker verification system. Four speakers' voices were used to train the recognition model, and fourteen (14) speakers' voices were recorded to test the model. Each speaker was made to say a word, and repeat same word with varied speech samples, which were compared with the database sample. The Euclidean distance was used as a comparative measure for identifying the original speaker. The performance rate of the system used in terms of authentication rate (AR), false authentication rate (FAR), and false rejection rate (FRR) at different Euclidean distance threshold values were also presented. Due to the effect of the variations in the speech samples for a given individual from the database sample, the system authentication rate drops from 100% to 45.45% at Euclidean distance threshold value of 6, from 100% to 50% at Euclidean distance threshold value of 5, from 100% to 50% at Euclidean distance threshold value of 4, and from 100% to 33.33% at Euclidean distance threshold value of 3.5..

Keywords: Speaker Recognition, Speaker Identification, Speaker Verification, Euclidean Distance, Voice biometric, Access control

1.0 Introduction

In this age of digital impersonation, biometric techniques are used as a check against identity theft. A person's voice is biometric [1]. And overtime, it has become a more reliable indicator of identity than legacy systems such as passwords and personal identification numbers (PINs) [2].

There are three basic ways to identify a person to a secured network. These can be based on; what you know, what you have, and who you are. However, each of these basic ways has their own advantages and disadvantages. The 'What you know' approaches such as passwords and PINs are not reliable since they can be lost, stolen, or guessed. 'What you have' technologies such as RFID cards and e-tokens can be stolen. But biometrics which is 'who you are' ensures a high level of security. The biometric method of recognition can be classified into behavioral and physiological approach. Speaker recognition is a behavioral approach of the biometric technique [1,2].

Speaker recognition is realized through feature extraction of the human speech and verification/identification of the speaker. Since speaker verification systems involve patterns which have some measure of variance across representative element of the speaker, it becomes a challenging problem [3]. In this case the speaker is often a variant in the database due to time, various recording conditions, and variable environmental conditions.

It is possible for a human to recognize a familiar voice, but with machines (e.g computer), it is more difficult. These difficulties are due to the fact that it is almost impossible for a phrase or word to be said exactly the same way on two occasions. And though it can be said the same way on those occasions, the recording alignment may not begin at precisely the same moment. Thus, the speaker verification/identification algorithm or system performance, irrespective of how good it may be, can be limited.

It is against this background that this paper presents a study to evaluate the performance of a generic speaker recognition system by varying the characteristics of the speech of an unknown speaker voice, and compare the resultant vector of the unknown speaker voice with database referenced speech samples using the Euclidean distance as a comparative metric.

Corresponding author: E-mail: joyokumo@yahoo.com, Tel.: +2348067147186

2.0 Background Study

Human speech is a complex signal, and this complexity is due to the large number of characteristics of the human speech which can be viewed on different levels; acoustic, linguistic, and psychological. Every person has a unique voice and when the same person speaks the same words on different occasions, the resulting sound may be not identical [4].

Speech signal is a slow varying signal (quasi-stationary). Its characteristics are stationary over a short period of time, but changes reflecting the different speech sound being spoken over a long period of time. Therefore the short-time spectral analysis is the most common way to characterize the speech signal [5].

Speech processing is a diverse field with many applications [6]. One of such applications is speaker recognition. Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals [7]. All speaker recognition systems contain two basic modules; feature extraction and feature matching. Speaker recognition encompasses verification and identification [6,8,9]. In speaker verification, the best match of an unknown speaker is identified from a list of known speakers, while speaker identification process decides if speaker is the person he claims to be [4,6,7,8,10,11,12].

Speaker recognition can be text-dependent or text-independent [3-7]. In text-dependent, utterances presented to the recognition system are known beforehand, and can be prompted. In text-independent case, no assumption about the text being spoken is made in which case the system must model the general underlying properties of the speaker's vocal spectrum.

In general, text-dependent systems like the system used for this study are more reliable and accurate since both the content and the voice can be compared [6,8]. In the text-dependent system, the claimant speak a phrase into a microphone, the signal is analyzed by a verification system, and appropriate decision to reject or accept the user's identity or possibly to report insufficient confidence and request additional input, is made. However, the choice of the technology (text-dependent or text-independent) to use in speech recognition system is application specific.

Due to the nature of the human speech, comparison between two vectors can be made. This make the Euclidean distance an accurate parameter to measure the closeness between the different speech frequency spectra (or coded vector) [13]. The Euclidean distance is the distance of criterion function [5,7]. The criterion function *E* is expressed as [5];

$$E=\sum_{i=1}^{k}\sum_{x\in Ci}|x-xi|^2 \quad (1)$$

The Euclidean distance is used to determine the nearest distance between each data object (unknown speech sample) and cluster center (database). Thus the Euclidean distance *d(xi,yi)* between one vector *x = (x1, x2, ……. xn)* and another vector *y = (y1,y2,…….yn)* is [5]

$$d(xi,yi)=[\sum_{i=1}^{n}(xi-yi)^2]^{1/2} \quad (2)$$

Where *x* and *y* represent the individuals.

In the speaker recognition stage, a distortion distance which is based on the minimum Euclidean distance is used when matching an unknown speaker with the speaker database [5].

3.0 Research Method

This work is aimed at realizing a Single-word ('close' or 'open') text-dependent using a generic voice authentication system. The work was realized in the MATLAB® 7.5 simulation environment. The input samples used in both training (enrolment) and the verification process were recorded in a mild noisy environment (lecture theatre). A built-in microphone of a Nokia 6120classic smart phone was used for recording. Voice samples were recorded at a rate of 8 kHz and 16bits quantization. The speaker was asked to say a text-dependent phrase 'close' or 'open'. There was no channel mismatch [6]. This microphone circuit and recorder GUI (Graphical User Interface) were used because it possessed optimized version of a microphone circuit with built-in filter circuitry for optimum performance during telephone calls. Sounds recorded were stored in the ".wav" windows sound file format, and imported into the MATLAB® 7.5 environment. This format can be read using the "***wavread()***"function.

The generic speaker recognition system used is VQ-based verification with speech pre-processing system. The system flowchart is shown in Fig.1. The speaker identification/verification algorithm was tested with different speakers' voices recorded and stored using Matlab software. Four (4) speakers' voices, two (2) male and two (2) female, both within the ages of 22 years, were used to train the VQ model. A total of fourteen (14) speakers' voices (5 females and 9 males) saying the same single-word one or more times, were recorded to test the model. The first voice samples of the authorized four persons were saved in a database, and subsequent voice samples were compared against all voice samples stored in the database. The Euclidean distance was used as a comparative metric for identifying the original speaker. The performance of the system used in terms of authentication rate (AR), false authentication rate (FAR), and false rejection rate (FRR) at different Euclidean distance threshold values were analyzed.

4.0 Results

The simulation results of the various voice samples matched with the database samples obtained in terms of Euclidean distance are presented in Tables 1. The inputs voice samples were coded as *XijW*. Where X is the male (m) or female (f) voice sound, *i = 1, 2, 3…..n* represent the first, second, ….. nth person used. *j =0, 1, 2, 3,……..n* represent the number of times the word was said with varied pitch. where *j = 0* represent the first time the word was said, *j* = 1 represent the second time the same i^{th} person said the word, and so on. *W* is the word 'Open' (o) or 'Close' (c) used in this study. Hence; f10c – first female saying the word close, f50c – fifth female saying the word close, f51c – fifth female repeating the word close with varying pitch for the first time, m93c – ninth male repeating the word close for the third time with varying pitch.

As seen in Table 1, the Euclidean distances recorded for each input signal against the authenticated user signals (f10, f50, m90, m20) are presented in rows. The minimum Euclidean distance in each row which is most likely the speaker's voice are highlighted as shown in Table 2.

5.0 Data Analysis

The Euclidean distance data obtained as in Table 1, and highlighted for the most likely speaker's voice, were analyzed with varying threshold values of the distances set at 6, 5, 4, and 3.5. These threshold values were useful to determine the discriminating ability of the system in terms of speaker voice authentication rate (AR), the false authentication rate (FAR), and the false rejection rate (FRR). The authentication rate (AR) which is the system's ability to make the right decision either to allow access or deny access is computed with equ.(3). The FAR which is a measure of the ability to identify the wrong person, and FRR which is a measure of the system's ability to reject the right person, were computed with equations (4) and (5).

$$AR = \mathbf{(}(TA + TR)/(TA + TR + FA)\mathbf{)} * 100 \quad (3)$$

$$FAR = \mathbf{(}(FA)/(TA + TR + FA)\mathbf{)} * 100 \quad (4)$$

$$FRR = \mathbf{(}(FR)/(TA + TR + FA)\mathbf{)} * 100 \quad (5)$$

Where TA is true authentication, TR – true rejection, FA - False authentication.

Based on the minimum Euclidean distance highlighted in Table 2, the values of AR, FAR, and FRR were computed for each of the samples (f10, f50, m90, and m20) stored in the database using equations (3) – (5). The results of these computations are shown in Tables 3 – 5 for different threshold values of the Euclidean distance.

6.0 Discussion

In this work, a VQ-based speaker verification model/system was trained with four (4) authenticated user voices which are coded as f10, f20, m90, and m20 to be able to have access control to a restricted environment. The VQ-based model is a text-dependent security system. The emphasis on this study is on determining the best match between an unknown speaker (test speech sample) identity from a list of known speakers (referenced samples) using the Euclidean distance parameter as a classifier or performance metric. Fig 1 is a model of the system used to test the speaker voice samples. The results obtained, after each simulation of the voice matching process in terms of the Euclidean distance are presented in Table 1. From a speaker voice match, the most likely speaker voice is coded as the smallest or the minimum Euclidean distance (see Table 2). The results in Table 2 were analyzed in terms of the system performance indicators (authentication rate (AR), false authentication rate (FAR), and false rejection rate (FRR)) with different Euclidean distance threshold values as presented in Tables 3 – 5. It was observed that as the characteristics of speech (pitch) from one test speech sample to the other against the referenced speech samples varied, the system performance indicators also varies. From Table 3-5, the system authentication rate drops from 100% to 45.45% at Euclidean distance threshold value of 6, from 100% to 50% at Euclidean distance threshold value of 5, from 100% to 50% at Euclidean distance threshold value of 4, and from 100% to 33.33% at Euclidean distance threshold value of 3.5. At Euclidean distance threshold values of 4 and 3.5, the system ability to discriminate between unknown and known speaker's speech was very high, hence achieving a 0% fake authentication. This means that the smaller the threshold value, the more sensitive the system is to slight variations in the characteristics of speech, even if it was the same speech sample from an authorized person. Thus, with a threshold value of 3.5, the system can only authenticate 33.33% of the speech samples tested under the same condition considered for this investigation.

7.0 Conclusion

In this work the performance of a speaker recognition system under varied speech characteristics using the Euclidean distance as a classifier is presented. The overall performance results obtained were assessed based on the Euclidean distance threshold values chosen. As the threshold value increases, the system ability to discriminate between imposters became poor. The Euclidean distance was found to be an efficient classifier in speech verification. However, variation in speech characteristics affected greatly the speaker recognition system performance, thus reducing the system's authentication rate.

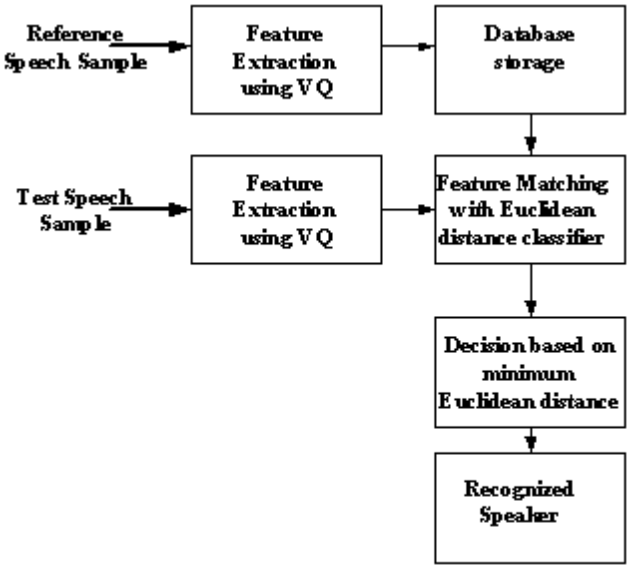

Fig. 1: A flowchart of VQ-based verification system

Table 1: Euclidean distance between various inputs and database templates

			output		
		f10c	f50c	m90c	m20c
	f10c	1.6495	5.6723	6.4611	6.6532
	f50c	5.9333	1.8072	7.2329	7.2384
	m90c	7.167	7.743	1.8278	5.6445
	m20c	6.6095	6.2449	5.3339	1.771
	f11c	4.1505	6.4021	8.3086	7.603
	f12c	3.5382	4.9923	5.9975	7.4288
	f20c	4.7144	5.5484	6.8875	6.3064
	f30c	8.6955	7.3797	10.1623	9.6799
	f40c	8.8659	6.9756	10.1099	7.3351
	f51c	4.5019	3.1312	5.5757	6.1154
input	m10c	6.7343	6.5188	6.4101	6.5953
	m30c	6.5409	5.0948	7.4864	7.8717
	m50c	7.0356	5.9555	5.2295	4.2823
	m60c	5.8391	5.0182	5.6846	4.9836
	m61c	6.9806	6.6783	7.1236	5.9848
	m62c	5.4983	5.1106	5.8045	4.766
	m70c	7.713	7.1311	9.695	8.1148
	m80c	7.8448	7.8268	9.0292	7.2629
	m81c	8.0727	7.4511	8.7358	5.9041
	m91c	7.1709	6.6393	3.957	5.3352
	m93c	7.2383	5.9227	3.9991	5.4189
	f13o	5.8698	6.1404	7.8232	8.4684
	f14o	6.9168	8.4927	7.1828	9.8251
	f41o	7.5523	6.7669	7.3356	6.1882
	f52o	5.6609	3.923	7.0047	6.6768
	f53o	6.1167	4.6383	7.5423	8.4038
	f54o	6.3337	4.3855	6.775	7.2109
	m11o	9.4182	8.2784	8.8254	10.2523
	m21o	6.2008	5.69	6.4454	5.648
	m31o	7.2256	6.774	6.9533	7.4157
	m40o	6.1302	5.2089	5.0767	4.8269
	m51o	7.6426	7.817	8.9706	6.6307
	m71o	6.5223	6.1093	8.3116	8.0582

Table 2: Minimum Euclidean distance.

			output		
		f10c	f50c	m90c	m20c
	f10c	1.6495	5.6723	6.4611	6.6532
	f50c	5.9333	1.8072	7.2329	7.2384
	m90c	7.167	7.743	1.8278	5.6445
	m20c	6.6095	6.2449	5.3339	1.771
	f11c	4.1505	6.4021	8.3086	7.603
	f12c	3.5382	4.9923	5.9975	7.4288
	f20c	4.7144	5.5484	6.8875	6.3064
	f30c	8.6955	7.3797	10.1623	9.6799
	f40c	8.8659	6.9756	10.1099	7.3351
	f51c	4.5019	3.1312	5.5757	6.1154
input	m10c	6.7343	6.5188	6.4101	6.5953
	m30c	6.5409	5.0948	7.4864	7.8717
	m50c	7.0356	5.9555	5.2295	4.2823
	m60c	5.8391	5.0182	5.6846	4.9836
	m61c	6.9806	6.6783	7.1236	5.9848
	m62c	5.4983	5.1106	5.8045	4.766
	m70c	7.713	7.1311	9.695	8.1148
	m80c	7.8448	7.8268	9.0292	7.2629
	m81c	8.0727	7.4511	8.7358	5.9041
	m91c	7.1709	6.6393	3.957	5.3352
	m93c	7.2383	5.9227	3.9991	5.4189
	f13o	5.8698	6.1404	7.8232	8.4684
	f14o	6.9168	8.4927	7.1828	9.8251
	f41o	7.5523	6.7669	7.3356	6.1882
	f52o	5.6609	3.923	7.0047	6.6768
	f53o	6.1167	4.6383	7.5423	8.4038
	f54o	6.3337	4.3855	6.775	7.2109
	m11o	9.4182	8.2784	8.8254	10.2523
	m21o	6.2008	5.69	6.4454	5.648
	m31o	7.2256	6.774	6.9533	7.4157
	m40o	6.1302	5.2089	5.0767	4.8269
	m51o	7.6426	7.817	8.9706	6.6307
	m71o	6.5223	6.1093	8.3116	8.0582

Table 3: Speaker Authentication Rate (%)

Euclidean distance threshold value / Database Speech Sample	6	5	4	3.5
f10	**66.67**	**50**	**50**	**33.33**
f50	**100**	**100**	**100**	**75**
m90	**100**	**100**	**100**	**50**
m20	45.45	54.55	90.9	90.9

Table 4: Failure Authentication Rate of the System (%)

Euclidean distance threshold value / Database Speech Sample	6	5	4	3.5
f10	**16.67**	**16.67**	**0**	**0**
f50	**0**	**0**	**0**	**0**
m90	**0**	**0**	**0**	**0**
m20	54.55	36.36	0	0

Table 5: Failure Rejection Rate of the System (%)

Euclidean distance threshold value / Database Speech Sample	6	5	4	3.5
f10	**16.67**	**33.33**	**50**	**66.66**
f50	**0**	**0**	**0**	**0**
m90	**0**	**0**	**0**	**0**
m20	**9.09**	**9.09**	**9.09**	**0**

References

[1] Pathak M, Portelo J, Raj B, Trancoso I, Privacy-Preserving Speaker Authentication, Springer-verlag, 2012, Berlin Heidelberg, pp 1 – 22.

[2] Sathish G, Saravanan S.V, Narmadha S, Maheswari S.U, High Confidence Hand Vein Pattern Authentication with an Improvement In Euclidean Distance Classifier, European Journal of Scientific Research, Vol.71, No.2, 2012, pp 243-254.

[3] Senturk A, Gurgen F.S, Feature Selection by Independent Component Analysis for Robust Speaker Verification, International Journal of Computer Science and Networking Security (IJCSNS), Vol.6, No.3B, March 2006, pp 229 – 239.

[4] Baltoi I.M, Toderean A.M, Sterca A, Ceptral-Based Speaker Recognition, Studia Univ. Babes-Bolyai Imformatica, Vol. LVII, No.2, 2012, pp 43 – 50.

[5] Kavita B, Shobak K, Speaker Recognition using K-mean Algorithm, International Technology Research Letters, Vol.1, Issue 1, 2012, pp 66 – 68.

[6] Campbell J.P, Speaker Recognition: A Tutorial, Proceedings of the IEEE, Vol.85, No.9, September 1997, pp 1437 – 1462.

[7] Kekre H.B, Kulkarni V, Performance Comparison of Automatic Speaker Recognition using Vector Quantization by LBG, KFCG, and KMCG, International Journal of Computer Science and Security (IJCSS), Vol. 4, Issue 6, pp 571 – 577.

[8] Reynolds D.A, An Overview of Automatic Speaker Recognition Research, ECTI Transactions on Computer and Information Technology, Vol.1, No.2, November 2005.

[9] Bimbot F, Bonastre J.F, Fredouille C, Gravier G, Magrin-Chagnolleau I, Meignier S, Merlin T, Ortega-Garcia J, Petrovska-Delacretaz D, Reynolds D.A, A Tutorial on Text-independent Speaker Verification, EURASIP, Journal of Applied Signal Process, Vol.2004, No.1, 2004, pp 430 – 451.

[10] Xie Y, Zheng X, A Speaker Verification System, Course Project for EEL6825 (Pattern Recognition), Instructor: Dr Clint Slatton, Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA, May 1, 2006, pp 1 – 14.

[11] Rabiner L, Juang B.H, Yegnanaraya B, Fundamental of Speech Recognition, Prentice-Hall, Englewood Cliffs, 2009.

[12] Furui S, 50 years of progress in Speech and Speaker Recognition research, ECTI Transactions on Computer and Information Technology, Vol.1, No.2, November 2005.

[13] Shanthini B, Swamynathan S, A Secured Authentication System for MANETs Using Voice and Fingerprint Biometrics, European Journal of Scientific Research, Vol. 59, No.4, 2011, pp 533 – 546