JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

SPEAKER VERIFICATION BY COMBINATION OF IMAGE AND SPEECH FEATURES USING ARTIFICIAL NEURAL NETWORKS

Jaweria Izhar[1], Dr. Satendra Kurariya[2]

[1]Research Scholar, Dept. of Computer Science, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road, Madhya Pradesh, India
[2]Research Guide, Dept. of Computer Science, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal Indore Road, Madhya Pradesh, India

Received: 19.01.2020 Revised: 05.02.2020 Accepted: 15.03.2020

ABSTRACT: Speech is the most normal path for human communication. It passes on a few kinds of data from the speech generation and recognition perspective. The speech signal passes on message and language data likewise called semantic data. Likewise, data about speaker's enthusiastic and physiological qualities can be gotten from speech signal. It additionally gives data about the earth in which the speech was delivered and the medium through which it was transmitted. Subsequently, speech signal conveys bunches of data which is encoded in a mind boggling structure. People can easily interpret the vast majority of this data. This has propelled specialists to create frameworks that consequently concentrate and procedure the immense measure of data in speech.

KEYWORDS: ANN, Communication, Speaker recognition.

 DOI: http://dx.doi.org/10.31838/jcr.07.04.300

I. INTRODUCTION

Speech is the most normal path for human communication. It passes on a few kinds of data from the speech generation and recognition perspective. The speech signal passes on message and language data likewise called semantic data. Likewise, data about speaker's enthusiastic and physiological qualities can be gotten from speech signal. It additionally gives data about the earth in which the speech was delivered and the medium through which it was transmitted. Subsequently, speech signal conveys bunches of data which is encoded in a mind boggling structure [1]. People can easily interpret the vast majority of this data. This has propelled specialists to create frameworks that consequently concentrate and procedure the immense measure of data in speech. The phrasing utilized for speaker recognition assignment is exhibited. In this manner, provokes identified with this present reality task of the speaker recognition innovation are talked about. It gives the issue definition and the fundamental thesis targets are recognized. The field of speech recognition started by taking a piece of information from the manner in which people speak with one another. In the comparable way, people can speak with machine through voice. There are different methods for speaking with machines, similar to console and different gadgets [2]. The significant disadvantage of these gadgets is that they are for the most part exceptionally moderate and unwieldy. Subsequently, connection of human with machine can be made simpler with speech. The primary goal of Automatic Speech Recognition (ASR) is to plan frameworks which can be worked by human directions/guidelines [3].

Speaker recognition is a dynamic biometric task. It originates from the more broad speech preparing territory. Comparative, to other speech-related recognition exercises (speech recognition, language recognition, and so on.) speaker recognition is a multidisciplinary issue. It is important to have comprehension of example recognition procedures and area information (Acoustics/Phonetics). The speaker recognition frameworks separate and describe speaker explicit properties from a given speech test so as to acquire the client's character. Speaker recognition is a wide territory that envelops speaker check, distinguishing proof, speaker division and ordering [4]. Speaker check framework attempts to confirm character guarantee of a client while the speaker distinguishing proof focuses to locate the best match from the speaker models present in the database.

A speaker recognition framework perceives people from their voice. The voice of each individual is unique. This is a result of the distinction in the states of vocal tract, size of larynx, and different organs engaged with voice

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

creation. Likewise, speakers vary in their style of talking or highlight, beat, style of pitch, method for articulation, decision of vocabulary and so forth [5]. Best in class speaker recognition frameworks utilize some of these highlights as one to accomplish increasingly precise recognition.

II. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) have risen as an incredible asset for individual validation strategy. The requirement for this system emerges at whatever point PC collaborates with this present reality. The Neural Network can understand the spatial, fleeting or whatever other relationship that can play out the undertaking, for example, grouping, forecast and capacity estimation (Choudhury et al., 1999). In this work, the fusion of the features extricated from the caught face images and speech articulations of the 50 speakers having a place with a similar age gatherings is made. These features are given to the PC recreated ANN model, which comprises of feed forward perceptron model and prepared by surely understood Back-Propagation Algorithm (BPA) [6]. The ANN model yield confirms the speaker characters in least number of preparing cycles and least info features of speech and image. This strategy is hearty and does not get influenced regardless of whether image is misshaped and speech signal comprises of commotion. By fusion method, the twisted images and boisterous signals of sound with less number of image and speech features given to information model of ANN gives 99% recognition score. The model for speaker verification is spoken to in Figure 1.

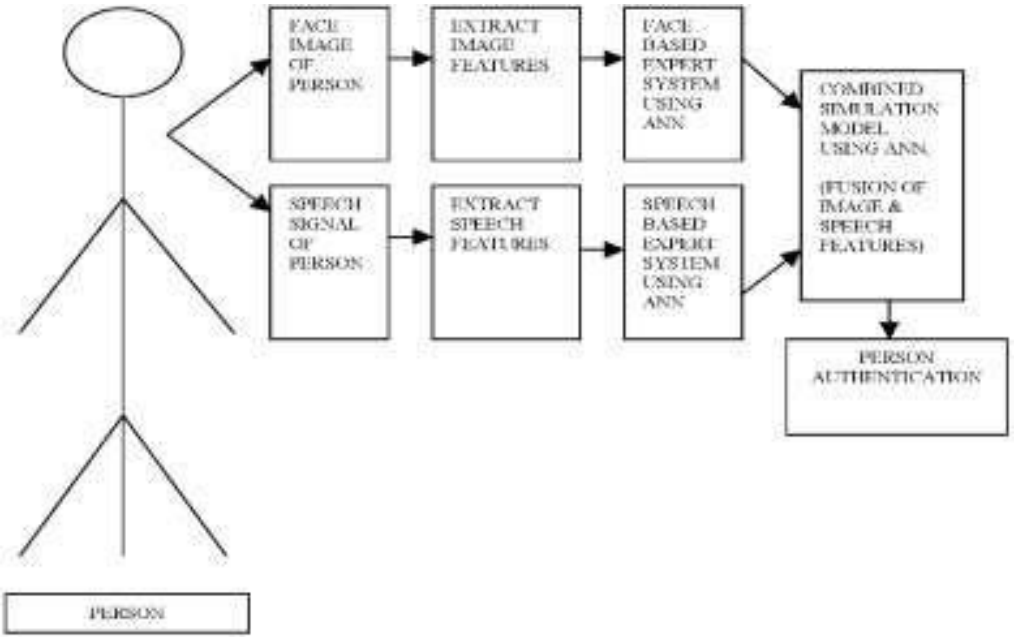


Figure 1: Model for Person Authentication

The perfect speech trademark for the speaker recognizable proof ought to happen normally and every now and again in the typical speech, store less information, utilize quicker or less mind boggling characterization systems, accomplish lower mistake, effectively quantifiable, ought not change over with time, ought to fluctuate however much as could reasonably be expected among the speakers yet be as predictable as workable for the specific speaker, and it ought not be influenced by the foundation clamor.

III.PROPOSED METHODOLOGY

The target of all Decision-Support Systems (DSS) is to make a model, which, when given a base measure of information/data, can create right decisions. Fusion of data is by all accounts worth applying regarding vulnerability decrease. Every one of the individual strategies creates a few blunders, referencing that the info data may be tainted and deficient [7]. Nonetheless, various techniques performing on various information should create various blunders, and accepting that every individual strategy perform well, blend of such multiple specialists ought to lessen generally order mistake and as an outcome underline right yields. The total plan for speaker confirmation utilized in this work of speaker verification by fusion of improved speech and image information utilizing ANN is appeared in the square outline.

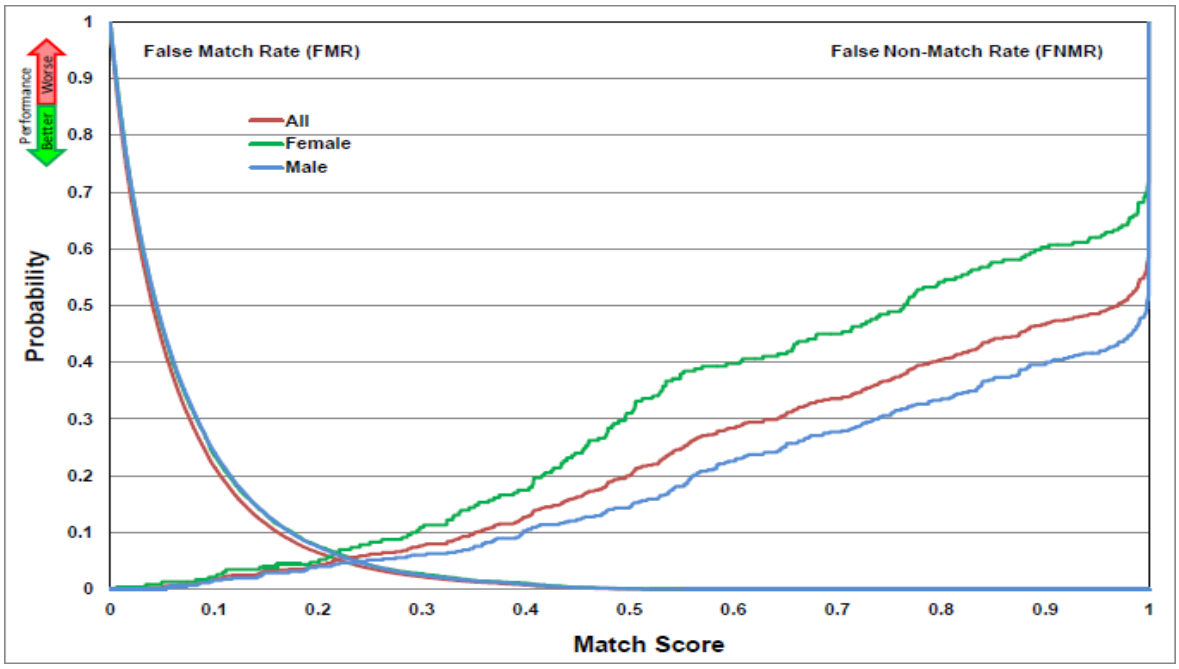
ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

The face recognition calculation utilized for enrolment and confirmation testing was Cognitec Face VACS SDK 8.9.5 [10], a well-known condition of-workmanship calculation utilized in face recognition applications for access control and cell phones. As far as speaker recognition, the audio tests utilized in the investigations were handled through Nuance Identifier ver. 9.4 system, a voice biometric system that verifies clients utilizing their own voice. Stage II of the Idiap Research Institute Mobio dataset [12] was chosen for the tests since it included between class biometric tests of speech and faces caught at the same time on a cell phone (Nokia N900). The audio-visual information gathered in this dataset approximates the nature of certifiable information gathered in non-controlled situations and incorporated an enormous enough number of characters for breaking down the outcomes with a high level of certainty. The Mobio dataset incorporates MPEG-4 audio video chronicles from 150 individuals (51 females and 99 guys) with 6 sessions for each individual and 11 accounts for each session. The accounts caught the members addressing short reaction questions, pre-characterized text read so anyone can hear and around 10 seconds of free speech. As featured by [12], because of the chronicles being caught utilizing a hand held gadget on various days and at multiple areas there is critical changeability in lighting, camera point and foundation of images just as the nature of speech audio. Furthermore an enormous variety in outward appearance, hairdo, garments, image sharpness and impediment were found in these images. For the trials, the Mobio dataset was separated up into enrolment and check testing sets. For every one of the 150 individuals, the longest account from the initial 5 sessions was chosen for testing, and the longest chronicle from the rest of the session was chosen for enrolment. The casing at the 3 second imprint was extricated to make the enrolment and check testing sets for face recognition. This brought about an enrolment database of 150 individuals (1 face image and 1 audio test for every individual) and a confirmation testing database of 150 individuals x 5 sessions (5 face images and 5 audio tests for each individual). Therefore, 750 certifiable and 55 875 impostor examinations can be gotten from this database. Be that as it may, because of the nature of the face imagery, a portion of the general population's images bombed the enrolment procedure by the Cognitec FR calculation and because of that images and audio records of these individuals were avoided from the investigation. So altogether, 676 certifiable and 48 594 impostor correlations were utilized to break down the system's exhibition. Table 1 underneath gives a breakdown of the examinations.

Table 1: Breakdown of the Comparisons for Genuines, Impostors and All Authenticities Versus Female to Female, Male to Male, Male to Female and All Genders

Authenticity	Female to Female	Male to Male	Male to Female	All
Genuines	229(3.9%)	447(2.0%)	0(0.0%)	676(1.4%)
Impostors	5634(96.1%)	21477(98.0%)	21483(100.0%)	48594(98.6%)
All	5863(11.9%)	21924(44.5%)	21483(43.6%)	49270(100.0%)

IV. RESULT AND DISCUSSION



ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

Figure 2: Cumulative Probability Plot – Face Recognition

The aggregate likelihood plots demonstrating the FMR and FNMR execution for the face and speech tests of the Mobio dataset are appeared in Figure 2 and Figure 3, separately. As a feature of the individual modalities execution appraisal, the dataset was additionally separated dependent on sexual orientation to build up any distinctions in execution between the two genders gatherings.

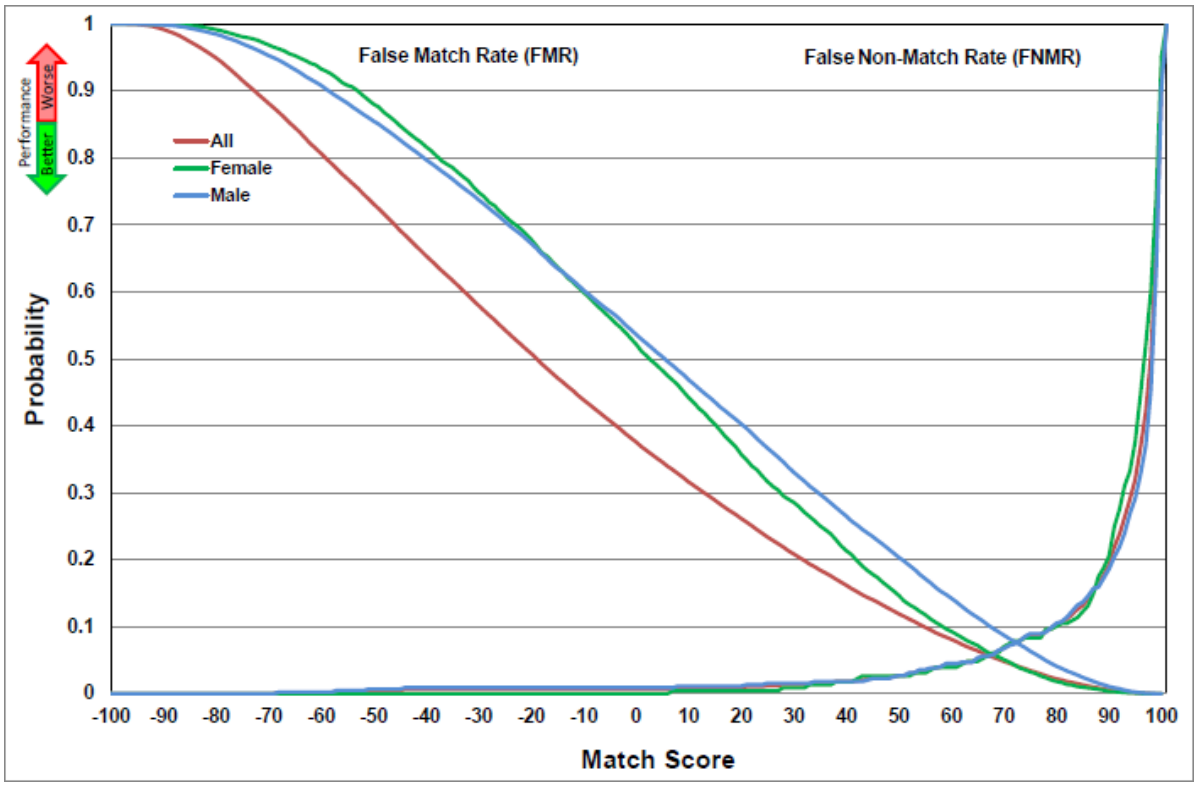


Figure 3: Cumulative Probability Plot – Speaker Recognition

For the right understanding of the total likelihood plot, the better FMR or FNMR execution is at the base of the plot and an insignificant cover among FMR and FNMR bends for the most part shows a superior performing methodology or dataset.

Utilizing these plots, an edge could likewise be set for every methodology (and its sex subgroup) with the end goal that both false match and false non-coordinate mistakes can be limited.

Regarding face recognition execution, Figure 2 demonstrates that there was a huge variety between the sexual orientation subgroups as far as FNMR with guys having lower FNMR, and, an insignificant variety as far as FMR. This suggests the FR calculation perceives guys more effectively than females and this is steady with the outcomes revealed in past examinations.

For speaker recognition, a little variety in FMR was found between the sexual orientation subgroups (as appeared in Figure 3), and a negligible variety as far as FNMR.

The FMR results demonstrate that guys with an improper case of personality may have a higher possibility of deluding the SR system if the system is set an edge ≥ 0. A superior FMR execution was gotten for the guys and females consolidated ("All") bunch since this imitates the zero exertion impostor situation where an individual makes no endeavor to build his/her possibility of achievement to cheat the SR system.

In this trial, an impostor could guarantee any personality, not simply those having a similar sexual orientation. Since the possibility of restoring a high match score when contrasting two voice tests of various sex is low, a superior FMR execution could be accomplished for the "All" gathering (as showed in Figure 3).

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

The general execution of the individual modalities and the impact of sexual orientation subgroups is likewise abridged in a DET plot (Figure 4), where the better performing methodology (or sex subgroup) is at the base left corner of the plot.

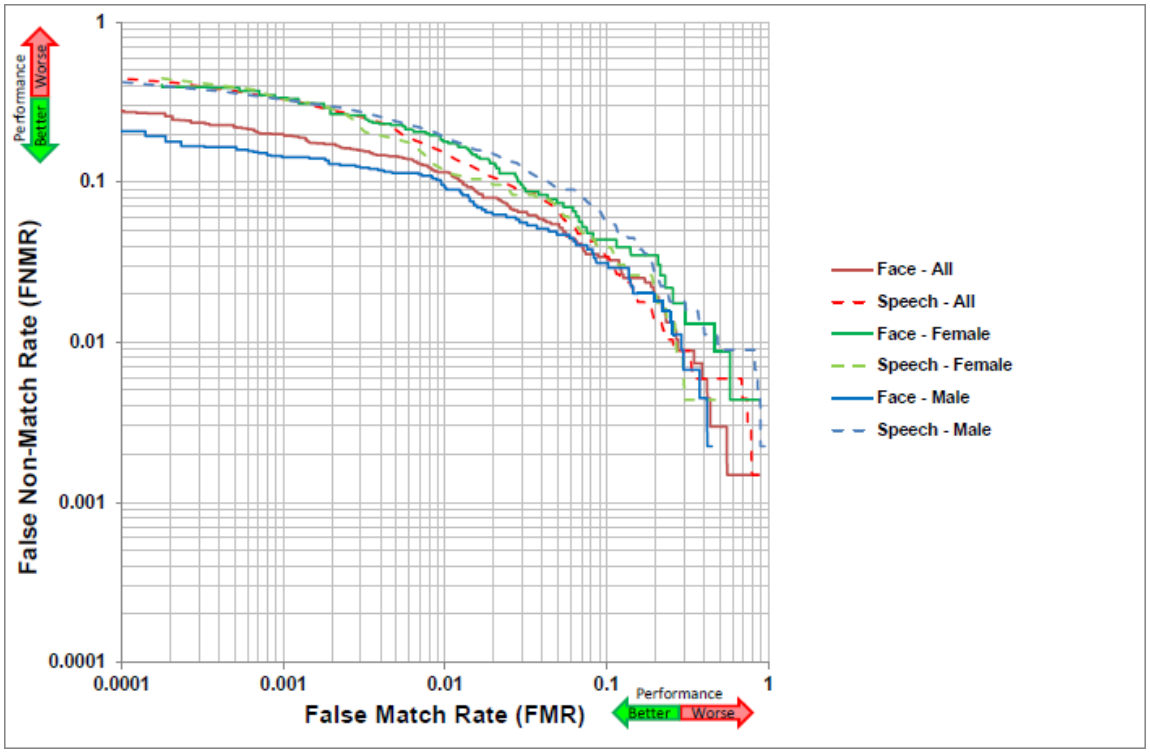


Figure 4: DET Plot – Face Recognition Vs Speaker Recognition

As appeared in Figure 3, the FR calculation seems to offer a superior by and large coordinating execution than the speaker recognition system at low FMRs. For the sex subgroups, the presentation seems, by all accounts, to be better when the face imagery is utilized as the biometric tests for the guys. For the female gathering, the presentation has all the earmarks of being better when their speech tests are utilized.

Multimodal fusion performance

In this area, the match scores from the face and speaker recognition systems were combined straight forwardly utilizing the weighted total strategy to show how the presentation of unimodal biometric systems can be improved. The fusion execution in this and resulting examinations were assessed by utilizing the zone under the Receiver Operating Characteristic (ROC) bend (meant as AUC).

For the weighted aggregate strategy, a pursuit is required to locate the ideal loads, that is, the loads that give the greatest AUC. Give w a chance to be the heaviness of the face scores, at that point the heaviness of the speaker scores is (10-w). In this investigation, we tried all w from 0 to 10 in additions of 0.1. For instance, when combining the face and speaker match scores the loads considered were (10, 0), (9.9, 0.1), (9.8, 0.2), … , (0.1, 9.9), (0, 10) separately.

Figure 4 demonstrates the weighted entirety fusion of face and speaker characters all in all dataset, just as, for every sexual orientation subgroup. The flat hub demonstrates the weight w of the speaker recognition system. At the point when w=0 the AUC is indistinguishable from the FR calculation alone, and is named on the left-hand vertical pivot. At the point when w=10 the AUC is indistinguishable from the speaker recognition system alone, and this is marked on the right- hand vertical pivot.

It tends to be seen from Figure 4 that for all gatherings, the general execution of the speaker recognition system is improved on the off chance that it is intertwined with the FR calculation, for all weightings analyzed. The ideal loads that give the greatest AUC (i.e., the best fusion execution acquired) appear to be the equivalent between the entire database and sexual orientation subgroups.

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

It ought to be noticed that when this dataset is inspected as a joined sexual orientation gathering, the fusion execution has all the earmarks of being less fortunate than utilizing the FR calculation alone if the weighting towards the speech match score is higher than 1.0. Notwithstanding, the fusion execution depends intensely on the decision of the standardization strategy and the order calculation utilized. These decisions are analyzed in the following area.

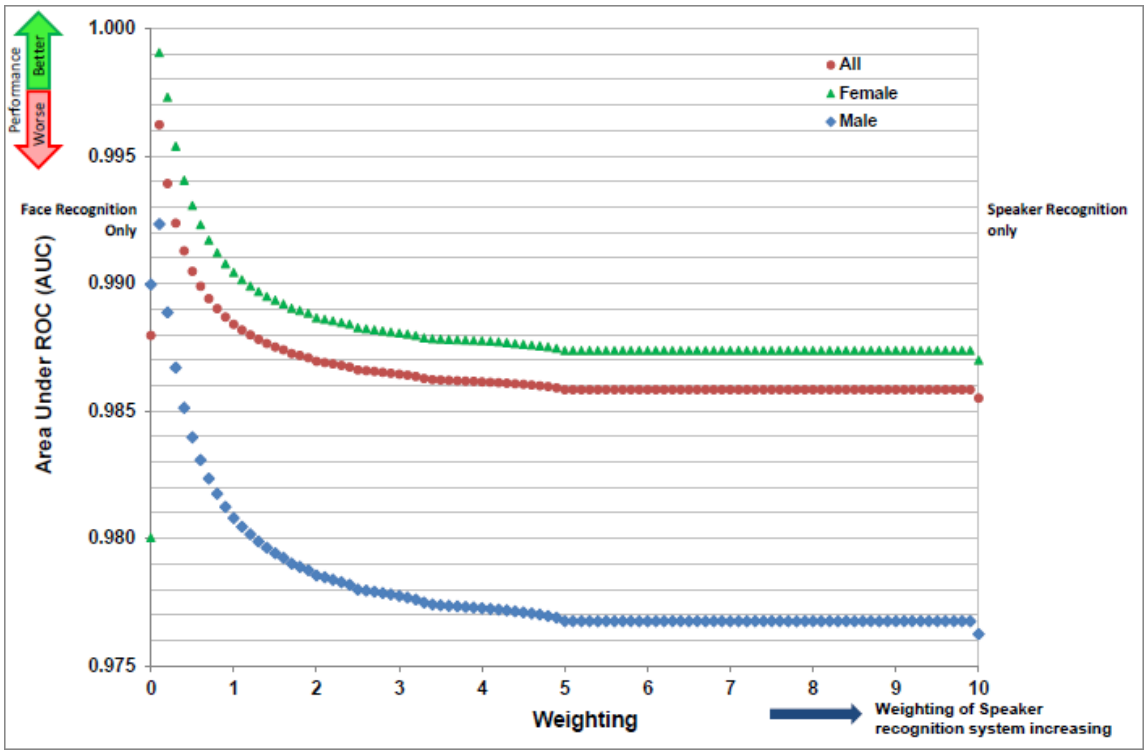


Figure 5: Weighted Sum Fusion of Face and Speaker Match Scores

V. CONCLUSION

The investigation results demonstrate that for same setup of ANN when the neural network is prepared for speech and image features taken together with less number of age as opposed to taking sound or image features independently. The quantities of features are additionally decreased i.e., for image we require just the separation among mouth and eyes and the separation between two eyes. From sound feature just the formants F1 to F3 alongside pinnacle and normal power other worldly thickness. Indeed, even a misshaped image of the speaker alongside couple of expressions of sound in uproarious condition can adequately validate the speaker.

The Performance of current speaker recognition systems is reasonable for some down to earth applications, for example, get to control, phone MasterCard's, and banking. The quantity of speaker verification applications will develop in the following five years and this will drive the exploration towards progressively vigorous modeling so as to cover however much as could reasonably be expected sudden commotion and acoustic occasions, and to decrease, (best case scenario) the quantity of innovative criminals. This work contributes another bearing of research in the field of image, speech and fusion strategy by ANN.

VI. REFERENCES

[1] Andreas Stolcke & Gerald Friedland,2010,„Leveraging Speaker Diarization for Meeting Recognition From Distant Microphones", *IEEE Int. Conference on Acoustics, Speech and Signal processing (ICASSP)*, pp. 4390-4393.

[2] Arlindo Veiga & Carla Lopes,2010,„Speaker Diarization Using Gaussian Mixture Turns and Segment Matching", *VI Jornadas en Technología del Habla and II Iberian SLTech Workshop, FALA* 2010, pp.409-412.

[3] Benjamin Bigot, 2010,„Detecting Individual Role using Features Extracted from Speaker Diarization Results", *Multimedia Tools and Applications*, vol.60, no.2, pp. 347-369.

[4] T.K. Ho, J.J. Hull, S.N. Srihari, Decision combination in multiple classifier systems, *IEEE Trans. Pattern Analysis and Machine Intelligence* 16 (1994) 66-75.

[5] Sanderson, K.K. Paliwal, Fast features for face authentication under illumination direction changes, *Pattern Recognition Letters 24* (14) (2003) 2409-2419.

[6] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification. *John Wiley & Sons,* USA, 2001.

[7] Mathieu Hu., 2015, „Speaker Change Detection and Speaker Diarization using Spatial Information", *ICASSP* 2015, vol.15, no.1, pp. 5743 – 5747.

[8] Lie Lu & Hong-Jiang Zhang, 2005, „Content Analysis for Audio Classification and Segmentation", *IEEE Transactions on Speech and Audio Processing,* vol.10, no.7, pp. 504-516.

[9] Konstantin Markov & Satoshi, 2007,"Never ending Learning system for on-line speaker diarization", *IEEE Workshop on Automatic Speech Recognition and Understanding,* pp. 699-704.

[10] Jourlin, J. Luettin, D. Genoud, H. Wassner, Integrating acoustic and labial information for speaker identification and verification, *In: Proc. 5th European Conf. Speech Communication and Technology, Rhodes, Greece,* 1997, Vol. 3, pp. 1603-1606.

[11] Hong, A. Jain, Integrating Faces and Fingerprints for Personal Identification, *IEEE Trans. Pattern Analysis and Machine Intelligence* 20 (1998) 1295-1306.

[12] Gerald Friedland,2012, „The ICSI RT-09 Speaker Diarization Systems", *IEEE Transactions on Audio, Speech and Language Processing,* vol. 20, no.2, pp. 371-381.

[13] Kumar, V.D.A., Kumar, V.D.A., Malathi, S, Vengatesan.K,Ramakrishnan.M "Facial Recognition System for Suspect Identification Using a Surveillance Camera" *Pattern Recognition and Image Analysis ,July 2018,* Volume 28, Issue 3, pp 410–420.

[14] Prabu, S., V. Balamurugan, and K. Vengatesan. "Design of cognitive image filters for suppression of noise level in medical images." *Measurement 141* (2019): 296-301