INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

System for Fusion of Face and Speech Modalities Using DTCWT+QFT and MFCC+RASTA Techniques

H C Shanthakumar[1], G S Nagaraja[2], Mustafa Basthikodi[3]*

1 Computer Science and Engineering, SJBIT, (Research Scholar, Jain University), Bengaluru, Karnataka, India
2 Computer Science and Engineering, RV College of Engineering (IEEE Senior Member), Bengaluru, Karnataka, India
3 Computer Science and Engineering, Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India

OPEN ACCESS

Received: 16.07.2021
Accepted: 19.11.2021
Published: 10.12.2021

Citation: Shanthakumar HC, Nagaraja GS, Basthikodi M (2021) System for Fusion of Face and Speech Modalities Using DTCWT+QFT and MFCC+RASTA Techniques. Indian Journal of Science and Technology 14(42): 3144-3156. https://doi.org/10.17485/IJST/v14i42.1316

* **Corresponding author**.

shanthkumarhc@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The main objective is to propose a multimodal biometric system by forming a fusion of Face and Speech modalities using DTCWT+QFT techniques for face and MFCC+RASTA Techniques for Speech recognitions. The experimental results are compared with existing works and analysed the performance with counterparts. **Methods**: The proposed model, make use of DTCWT and QFT techniques to extract the features of face images and perform fusion of both. The MFCC and RASTA techniques are implemented to extract features of speech data and then fusion is applied. Various databases discussed and utilized for both face and speech recognition system proposed. **Findings**: The results of experimentation are compared with existing systems and analysis proved than the proposed system is placed in better position. The fusion of DTCWT and QFT techniques for face recognition system is implemented and the results using performance parameters such as False Acceptation Ratio (FAR), False Rejection Ratio (FRR), Total Success Rate (TSR), Partial Error Rate (PER), Equal Error Rate (EER) are tabulated for six different types of face data sets. The average performance of the results is compared with four existing fusion techniques and showed that the proposed system performs better. The fusion of MFCC and RASTA techniques for speech recognition system is implemented and the performance is measured by calculating accuracy, precision, recall and F1-score. These results are compared with five different schemes and proved that proposed system of fusion of face and speech traits works better for human recognitions. **Novelty**: Fusion of two algorithms for face recognition is implemented and the results analysed. Then the fusion of two algorithms for speech recognition is implemented and the results are analysed. The novel approach is presented to combine both face and speech recognition system in to single system to improve the security using multimodal biometrics.

Keywords: DTCWT; QFT; RASTA; MFCC; Feature Extraction; Fusion

1 Introduction

The advancements in biometric systems using various modalities recognize a particular human on behavioural and physiological traits in faster and efficient manner. The quantity of studies in regards to recognition systems for face and speech modalities have been expanding every year. People can comprehend and imagine different emotions consistently. This should be possible by seeing different elements like movements of facial muscles, voice, hand signals, and so on [1]. The recognition framework for speech has been utilized in numerous areas. The speech or voice recognition system is a step-to-step process of distinguishing words or sentences by means of a machine. This process is in need of exact algorithms, such as classification and feature extraction algorithms. The effective strategies for feature extraction these days included are RASTA (Relative Spectral Filtering) and MFCC (Mel frequency cepstral coefficients). The MFCC is one of well-known techniques for feature extraction because of the better accuracy it has. Few of the studies concerning voice recognition framework making use of MFCC as feature extraction technique are conducted by many of the researchers [2,3]. The speech enhancement technique RASTA was basically emerged with objective of subsiding addictive and unwanted disturbances in Automatic voice recognition systems. The technique RASTA not just eases the effect of noise in voice signal however it likewise upgrades the quality of voice with background disturbances. Consequently, RASTA is technique of modulation frequency band sifting which could be utilized either in log spectral domain or cepstral domain, where RASTA filter band goes through every coefficient of features. The procedure of RASTA speech processing technique, is illustrated in the block diagram given in Figure 1. The more standardized mechanism MFCC is dependent on human ear's known variation of frequency with critical bandwidth. MFCC is advantageous in decreasing the recurrence data of the input voice signal into coefficients, it is a quick, dependable and simple computation method. The primary goal of Mel-Frequency cepstral coefficient is to impersonate human hear-able framework and thus is utilized for processing of speech. The procedure of MFCC technique for feature processing, is illustrated in the block diagram given in Figure 2.

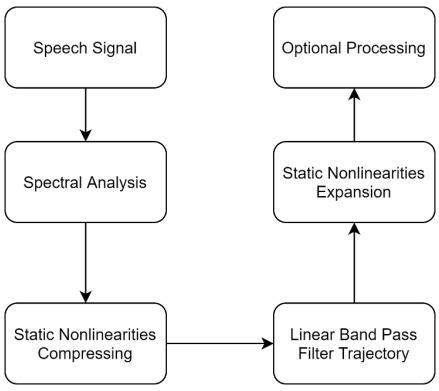

Fig 1. Diagrammatical Representation of RASTA technique

The face biometric modality is mostly used and powerful among all types of biometric traits as samples of face pictures are gained utilizing nonintrusive technique and with no collaboration of an individual. The technique DTCWT (Dual Tree Complex Wavelet Transform) is a novel enhancement strategy of DWT. It's anything but viable

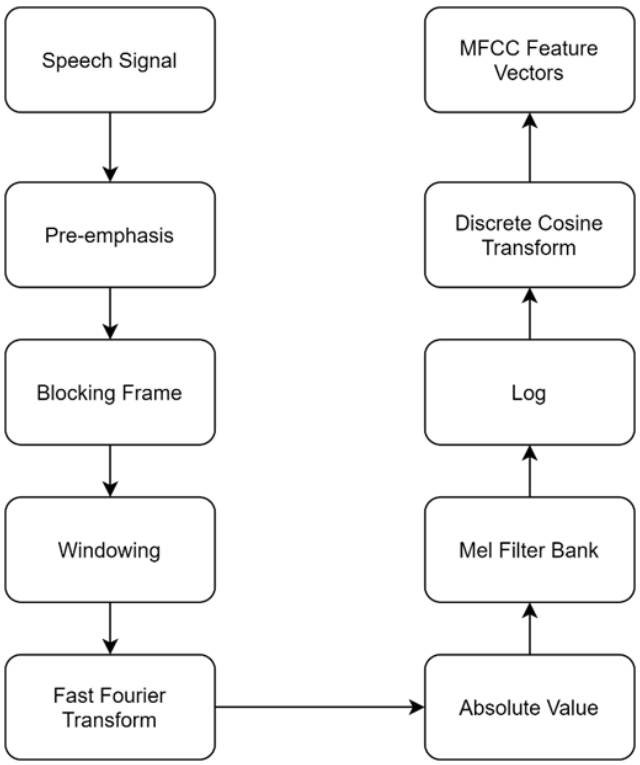

Fig 2. Diagrammatical Representation of MFCC technique

strategy for implementing an analytical wavelet transform. The complicated coefficients created by DTCWT presents restricted excess and permits the change in providing directional selectivity and shift invariance of filters.

The technique DTCWT could be carried out making parallel use of 2D divisible two real wavelets transform. The principal real wavelet transform can be implemented making use of high pass and low pass filter coefficients $H_0(k)$ and $H_1(k)$ applied along the row and column dimensions of 2D information that makes structure of DTCWT with Upper Filter bank.

The second Real wavelet transform that describes Lower Filter bank of DTCWT could be developed by making use of high pass and low pass filter coefficients G_0 (k) and $G_1(k)$ that are around logical to coefficients of upper filter bank bringing the results in ideal reproduction of incoming information of images. The QFT (Quick Fourier Transform) is a faster calculation algorithm for Discrete Fourier Transform utilized in applications of signal processing such as Correlation analysis, Linear Filtering and spectrum analysis which includes higher time for computation, that results in moderate efficient algorithms. In QFT, the sequence of data is divided into smaller sequences until we are able to get sequences of single-point. Considering N = 2s, such decompositions can be computed s = $\log_2$ N intervals. Hence, the total count of complicated multiplications decreased to (N/2) $\log_2$ N versus N2 complicated multiplication of straight forward calculation of DFT. In same way, the count of complicated additions is decreased to N $\log_2$ N considered to N2 – N complicated additions of directly DFT calculation.

There is good amount of work carried out in the field of speech and face recognition systems by various researchers. For the detection of active speaker, the works done in the paper [(4)] presents a procedure for effective fusion of correlated auditory and visual information for active speaker detection. The redundant data, noise information produced during the time spent single-modular component extraction, and conventional learning algorithms are hard to acquire ideal performance of recognitions. The authors in [(5)] propose a deep learning based multimodal fusion of emotion-based recognition strategy for voice expressions. During the process of person's social and day to day activities, voice, text and expressions of face are considered as primary channels to pass on human feelings. In this work [(6)], based on voice, motion ant text, a fusion strategy for multi-modal emotion recognition is proposed. The investigation of a robust strategy for multimodal emotion detection when a conversation is happening, is presented in the works [(7)]. Three distinct models for text, video and audio modalities are fine-tuned and organized on MELD.

The work in article [(8)] gives a careful assessment of the various studies that have been directed since 2006, when techniques such as deep learning initially emerged as another space of Machine learning, for speech applications. The work in article [(9)]

depict an execution of speech identification to pick and place an item utilizing Robot Arm. To get the component extraction of speech signal utilized Mel-Frequency Cepstrum Coefficients (MFCC) strategy and to gain proficiency with the data set for recognition of speech made use of Support Vector Machine (SVM) technique, the algorithm dependent on Python. The authors in article [10] presented methodology for concluding the quality of descriptions of grammar of Tatar language and lexicon's degree of coverage.

The historical activities of technologies for recognition of face modality, the present status of-the-art techniques, and directions for the future are discussed [11]. This explicitly focus on the latest data sets, 2D and 3D face identification techniques. In addition, this gives specific consideration to deep learning method as it describes the fact in those fields. The face ID making use of DTCWT has been utilized adequately for database L-Spacek [12]. The pre-handling is refined on face picture for acquiring uniform size for every one of the pictures and Dual-Tree Complex Wavelet Transform is utilized in the resized picture of appearances for getting features of DTCWT and these attributes are considered as the last ones. The Euclidean Distance is adjusted for coordinating. The authors [11], propose an ASR made with CNN where the exhibition of two element extraction techniques, to be specific Mel Frequency Cepstral Coefficients (MFCC) and Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) are looked at on Bangla's disengaged words comprising digits and voice commands. This work also adds to the writing of Bangla ASR using few unique approaches. First and foremost, Effects of commotion is probed Bangla voice commands just as secluded words in CNN based ASR. Furthermore, the exhibition of MFCC and RASTA-PLP are analysed in disturbing environment utilizing CNN based classifier.

The acoustic extraction approach improvement dependent on a hybrid procedure comprising of Perceptual Wavelet Packet (PWP) and Mel Frequency Cepstral Coefficients (MFCC) is presented [13]. The exhibition of wavelet change can be utilized as a denoising strategy in voice identification framework utilizing MFCC technique for feature extraction is done [14]. The good variables to be utilized in the framework are Minimax soft thresholding determination rule on level 10 for the decomposition. The authors in [15], worked on a structure modification for the power normalized cepstral coefficient (PNCC) framework for separating more relevant features of speech, explicitly at low sign to noise proportions (SNRs), without influencing the framework execution for undistorted speech. The article [16] significantly zeroing in on the improvement of speech acknowledgment by using the hybrid features such as MFCC and LPC, individually. By utilizing the spectral deduction strategy in pre-preparing stage, they were successfully taken out the noise from the voice with the equipped for viable extraction of source voice from noisy surroundings. The extraction of features has performed by LPC and MFCC strategy precisely with each feature types involving echo-based varieties of phases. The automatic implementation of the multimodalities could be achieved using parallel algorithms [17] in multicore systems. In the work proposed by authors [18], recommended a visually impaired multimodal watermarking strategy for biometric confirmation frameworks. The introduced approach depends on fusing face and finger impression modalities by means of the mix of DTCWT and DCT recurrence space methods to get and improve the presentation of the biometric verification frameworks. The directed tests showed the capacity to validate images with an impressively high precision level. The authors in [19] [20] presented an approach which used the face detection system as a biometric technology to mark students and employee's attendance in the organizations. The researchers in [21] and [22] presented an overview about multimodal biometrics by making use of face and ear. The work in paper [23] proposed a hybrid methodology by joining cascading and fusion of multimodal biometrics system making use of face and fingerprint traits.

2 Face and Speech Datasets

The various available and created data sets used for face and speech recognitions are discussed in this section.

Spacek Face database: This database created by Libor Spacek [24] includes data of 395 individuals with 20 images per individual. There are 7900 images of both male and female genders of various racial origins. The Figure 3 shows face image samples of Spacek data sets.

Extended Yale Face Database B +: Extended Yale Face Database B+ [25] contains 16128 pictures of 28 human category with 9 poses and 64 illumination constraints. The format of data of this particular database is the similar to the Yale Face Database B. The Figure 4 gives the face image samples of Extended Yale Face Database B (B+).

Near Infrared Face Database: The database [26] contains varieties of expression, poses, scale, illuminations, blurring and combinations of all of them. The database contains 115 humans and 15 pictures of every person. This standard database contains both male and female images with and without spectacles. The face images are of JPEG format with each image of size 768*576. The selected sample of Near infrared face pictures are shown in Figure 5.

ORL (Olivetti Research Lab) database: The face database ORL [27] consists of 400 pictures of size 112 x 92. This includes pictures of 40 persons, and 10 pictures for every person. The pictures were captured at numerous times, facial expressions and lighting. The faces are in a position of upright in frontal view, along with a slight left-right rotations. The selected sample of ORL face pictures are given in Figure 6.

Fig 3. Selected Sample of L- Spacekface pictures of human

Fig 4. Samples of Extended Yale Face Database B+

Fig 5. Selected sample of Near Infrared face pictures

Fig 6. Selected sample of ORL face pictures of human

Data sets for speech recognition system, the LibriSpeech corpus is used [28], that is taken from audiosets and consists of English speech of 1000 hours sampled at 16 kHz. As per the texts in the audiobook, the speakers recorded their sentences. Because of the low error rate, the database is best suitable for training and evaluation of speech recognition systems. The LibriSpeech database made use in our research contains the voice belong to 16 speakers, among those, 8 are men and 8 are women, with each speaker speaking 11 distinct sentences.

3 Methodology

In this section, proposed methodology for face and speech recognition system is discussed. The extracted features of QFT and DTCWT are fused to generate final face feature set. The extracted feature of MFCC and RASTA are fused to get final results as final speech feature set. The model proposed concentrate on enhancing the recognition rates for both face and speech modalities. The diagram illustrating the proposed model is given in Figure 7. The face and speech databases mentioned in the previous section are used in the proposed model the extract required images and voice inputs to perform the pre-processing and for the analysis of performance.

The various face databases contain numerous dimensions in the face; therefore, the images may be processed into uniform sized images. Every image is processed to 2p × 2q where p and q are integer variables. The images of face are processed to size of 128 × 512. The algorithms DTCWT and QFT are applied to those images which are resized.

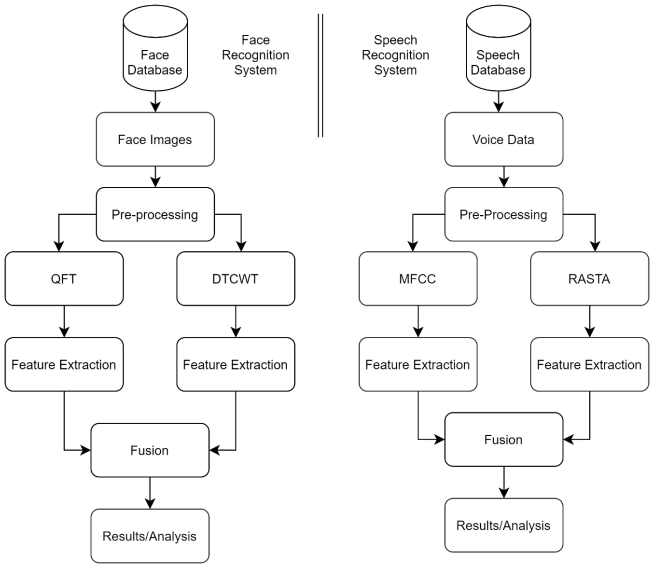

Fig 7. Proposed Methodology Block diagram for Face and Speech recognition

The Feature Extraction of Two Dimensional DTCWT follows the steps given below:

Firstly, an image given as input is made to decompose using two-dimensional DWT. Our model proposed applies five stage DTCWT on images of face, that gives sixteen sub levels at every stage, four sub levels having lower frequencies and twelve sub levels having higher frequencies. In each stage, size of the image is decreased to 50% of original size of the image, that is, in to 4 x 16 image size.

Secondly, each two respective sub levels that contain the similar pass levels are linearly joined by differencing and averaging. Resultantly, the sub levels of two-dimensional CWT in every stage are computed as (SPx+SPy)/$\sqrt{2}$, (SPx − SPy)/ $\sqrt{2}$, (PSx +PSy)/$\sqrt{2}$, (PPx − PPy)/$\sqrt{2}$, (PPx + PPy)/$\sqrt{2}$.

Thirdly, For the recognitions of the face features, magnitudes of real levels and imaginary levels are considered. Two-dimensional DWT on every decomposition obtains three higher frequency levels such as PS, SP and PP that provides a directional data. The DTCWT developed making use of two-dimensional real wavelet transform generates six complicated wavelets generating directional data on various directions by considering real part and imaginary part of every complicated wavelet. The magnitude of a set of six complicated wavelets are computed by the equations given below, Equations 1 and 2 and the finalized magnitude coefficients are produces by applying the concatenation operation as given in the Equation 3.

$$x_{pq} = \text{sqrt}\left(x_p^2 + x_q^2\right) \tag{1}$$

$$x_{yz} = \text{sgrt}\left(x_y{}^2 + x_{\bar{z}}{}^2\right) \tag{2}$$

$$\text{X} = \left[\ x_{pq} + \quad x_{yz}\ \right] \tag{3}$$

In these equations, x_p, x_q, x_y, x_z arerespective to DTCWT coefficient vectors of higher frequency having size 1 × 192 in 5-Stage DTCWT. The resultant vector X of features 1 x 384 generated by applying the concatenation operation to the magnitudes x_{pq} and x_{yz}.

For the Quick Fourier Transform (QFT) feature extraction, apply 2-D QFT on the Pre-processed images of face having size 128 × 512 in order to obtain coefficients of QFT by making use of the Equation 4. The absolute values of QFT coefficient are sorted in nonincreasing order. Highly dominant 384 coefficients are decided to consider as features of QFT. Dominant features of QFT are fused with DTCWT features by applying arithmetic addition operation to obtain resultant features, to get improved recognition rate for a person. The resultant features of fusion are generated making use of the Equation 5. In order to get the features of test images, the QFT and a five-stage DTCWT are computed on test images. These features are analysed with features of database pictures applying Euclidean distance given in Equation. 6, to produce values of FAR, FRR, TSR and EER.

$$F(p,q) = \sum_{k=0}^{n-1} \sum_{i=0}^{m-1} f(k,i)\, e^{-j2\pi\left(\frac{pk}{n} + \frac{qi}{m}\right)} \tag{4}$$

In the above equation, F(p,q) represent QFT Coefficient and f(k,i) represent input face image.

$$The\ resultant\ feature = \sum_{n=0}^{384} (FFT_n + DTCWT_n) \tag{5}$$

The FFT_n and $DTCWT_n$ provides coefficients of Quick Fourier Transform and DTC wavelets.

$$\mathrm{ED}(x,y) = \mathrm{sqrt}\left(\sum_k^n (x_k - y_k)\right)^2 \tag{6}$$

In Eqn.6, the x_k represents feature value for images from database, and y_k represents feature value for test images.

The voice data from the speech database are pre-processed, before we apply the MFCC and RASTA techniques for the feature extractions.

The scale of Mel-Frequency is a lower frequency which is linear within 1000 Hz and logarithmic higher frequency more than 1000 Hz. Equation 7 provides the relationship of Mel scales to frequency in Hz.

$$\mathrm{FRmel} = \begin{cases} 2595 * [\log] 10 \left(1 + \dfrac{\mathrm{FRhz}}{700}\right), & FRhz > 1000 \\ FRhz, & FRhz < 1000 \end{cases} \tag{7}$$

In above Equation 7, FRmel is Mel scale and f is frequency in Hz. The procedure in Mel scale to the frequency spectrum with the working function of ear of person as a filter is via Filter Bank. Suppose F[N] spectrum is input for the process, and then output is M[N] spectrum that is the F[N] modified spectrum which has Power Output of those filters. The spectrum coefficient of Mel is specially determined to be 20. With respect to the Cepstrum, the persons listen to speech information depending on time domain signals. In this particular phase, Mel-spectrum would be transformed into time domain by making use of Discrete Cosine Transform (DCT). And the outcome would be MFCC. The cosine transformations expressed using the equation given in Equation 8.

$$Xi = \sum_{i=1}^{N} Zi\ cos(i(j-1)/2\ \frac{Pi}{j} \tag{8}$$

The computation at Equation 8, provide X_i as the coefficient of MFCC, Z_i is Mel frequency power spectrum, i = 1, 2, 3, …, N, where N is the count of coefficients desired and M is presented as count of filters.

The specialized level-pass filter joined to each frequency sub-levels in MFCC algorithm to make it smooth out shorter noise variations and to reduce any constant disturbances in the speech channel. As shown in the Figure 7 the voice data is pre-processed and given to RASTA mechanism for extraction of relevant features in separate channel. The fusion of MFCC-RASTA will make a good matching result and in turn to increase in the recognition rates.

4 Results and Discussion

The experimentation of face recognition system in proposed model is done using MATLAB. The various set of face images are used from the databases discussed in section-III, such as Spacek Face database, Extended Yale Face Database B+, Near Infrared Face Database and ORL database along with the Indian male and Indian female databases.

The parameters for the performance analysis, experimental results by making use of DTCWT, QFT, and fusion of these techniques are analysed. The numerous combinations of Human inside database (HID) and Human outside database (HOD) of each database is used to know the variations in the performance parameters.

The performance parameters such as False Acceptation Ratio (FAR), False Rejection Ratio (FRR), Total Success Rate (TSR), Partial Error Rate (PER), Equal Error Rate (EER) are used for evaluations are defined below:

Let A be the Count of human faces accepted in the outside database and B be the total count of humans available outside database. Then,

$$FAR = \frac{A}{B} \tag{9}$$

Let C be the count of genuine humans rejected in the inside database and D be the total count of humans in database. Then

$$FRR = \frac{C}{D} \tag{10}$$

Let X be the count of matched humans and Y be the total count of humans available inside database. Then,

$$TSR = \frac{X}{Y} \tag{11}$$

$$HER = \frac{(FRR + FAR)}{2} \tag{12}$$

The EER describes rate error, where FRR and FAR both are equal. These parameter values of DTCWT, QFT and fusion of these mechanisms are recorded in Table 1 , with various combinations of HID and HOD values.

Table 1. Comparison of Performance of proposed DCTWT and QFT techniques

Database	HID: HOD	DCTWT			QFT			FUSION		
		TSR (%)	EER (%)	Max TSR (%)	TSR (%)	EER (%)	Max TSR (%)	TSR (%)	EER (%)	Max TSR (%)
Spacek	20:30	85	15	100	80	20	100	82.5	18	100
	30:20	80	20	100	80	20	94	80	18	100
	10:30	88	12	100	84	16	100	88	14	100
	30:10	92	6	98	88	10	92	92	10	100
Extended Yale	12:8	84	14	98	86	12	98	85	15	100
	8:12	90	10	96	82	16	96	92	10	100
	15:20	91.43	8.5	99	85.71	15	100	92	12	100
	20:15	88.57	10.5	98	82.86	17	98	98.8	9	100
Near Infrared	10:20	88	12	100	85	15	100	86	14	100
	20:10	85	15	100	88	12	100	86	14	100
	15:20	85.71	14	100	82.86	16	98	88	12	100
	20:15	91.43	8.5	100	85.71	15	98	90	12	100
ORL	30:10	88	12	100	92	10	98	90	12	100
	10:30	84	16	95	88	12	100	88	14	100
	20:30	80	21	92	85	15	94	85	15	100
	30:20	80	21	92	85	16	92	84	16	100
Indian male	10:15	82.5	17	98	85	15	98	96	8	94
	15:10	82.5	17	98	85	15	98	94	10	94
	20:15	87.5	13	96	83	15	95	84	15	92
	15:20	87.5	13	96	83	15	95	88	12	92
Indian female	15:18	90.90	10	93	87.88	13	92	98.5	8	92
	18:15	90.90	10	93	87.88	13	92	92	12	93
	20:25	86.66	14	95	80	16	90	88	12	88
	25:20	88.88	12	95	80	16	90	92	15	90

The Table 1 gives the recorded values of optimum TSR percentage, EER percentage and percentage of maximum TSR for DTCWT, QFT and fusion of these techniques, which are calculated against six different databases mentioned in the table. The databases such as Spacek, Extended Yale, Near Infrared and ORL records maximum TSR value 100, whereas Indian male database records maximum TSR value 98 and Indian female records 95 maximum TSR. The EER is recorded 6, 8.5, 10, 10 and 8 for Spacek, Extended Yale, Near Infrared, ORL, Indian Male and Indian female databases respectively. The performance evaluation of proposed fusion technique for face recognition is done with various existing techniques proposed by different researchers.

Table 2. Results of comparisons of proposed technique with related existing techniques

Techniques	Spacek		Extended Yale		Near Infrared		ORL		Indian male		Indian female	
	TSR%	EER%	TSR%	EER%	TSR%	EER%	TSR%	EER%	TSR%	EER%	TSR%	EER%
FFT + DWT [27]	90	10	88.4	10	72	19	90	10	84.40	16.40	87.61	8.21
LBP +DWT +SOM [28]	60	30	64	36	76	34	60	30	55	40	72	30
DWT +SVM [29]	88	12	65	36	84	15	90	10	58.33	44	77.5	12.5
RLM for Canny Edge [30]	60	40	70	32	82	18	50	50	70	32	80	19
Proposed	92	6	98.8	8.5	91.43	8.5	92	10	96	8	98.5	8

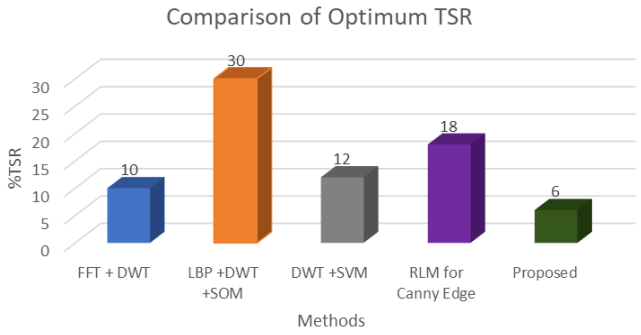

Fig 8. Graphical Analysis of Total Success rate

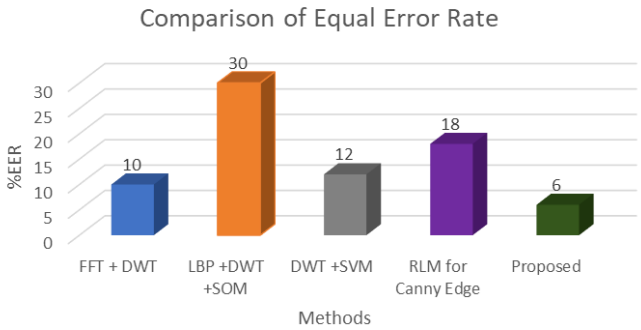

Fig 9. Graphical Analysis of Equal Error rate

The Table 2 gives the results recorded towards performance comparison of proposed fusion techniques with the existing mechanisms presented by researchers in [29], [30], [31] and [32]. The percentage values of optimum TSR are higher and percentage

of Equal Error Rate values are less in proposed technique when compared with available techniques. The graphical illustration of Total Success Rate and Equal Error Rates are demonstrated in Figures 8 and 9. The rates of recognition and Half error of proposed technique is compared with the TLPP and BSIF+TLPP(7x7), FFT+DWT techniques presented in the work[30] and[29]. The graphical analysis given in Figures 10 and 11 shows that the proposed technique's performance is better placed.

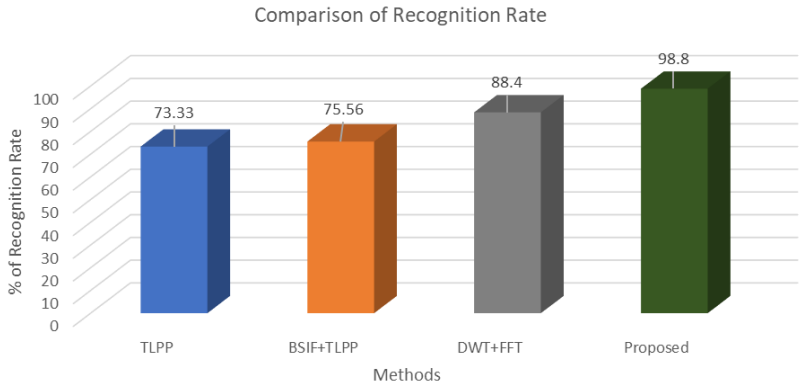

Fig 10. Graphical Analysis of Recognition Rates

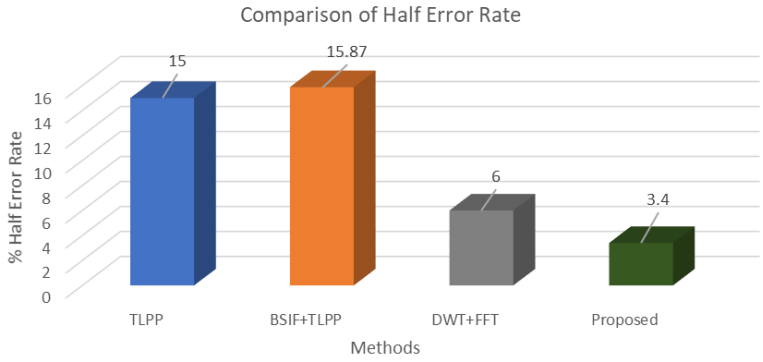

Fig 11. Graphical Analysis of Half Error Rates

The speech recognition system making fusion of the algorithms MFCC and RASTA are evaluated for the performance against the data sets LibriSpeech corpus, which consists of English speech of 1000 hours, with 16 speakers containing both men and women, speaking distinct words in their sentences. The results are compared with existing speech recognition methods. While doing the experimentation the recognition rate is computed by taking in to account the total count of speakers and count of right matches. The performance metrics including accuracy, F1-score, recall and precision are utilized for the evaluation of performance. Where, Accuracy (ACC) is count of data matched rightly out of total data sets, Recall (RC) gives the speech proportion checked positive and recognized, Precision (PR) gives the speech proportion more precisely recognised and F1-Score is computed from recall and precision values. The parameters such as True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) are made use for computing these metrics are defined below.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

$$RC = \frac{TP}{TP + FN} \tag{14}$$

$$PR = \frac{TP}{TP + FP} \tag{15}$$

$$F1 - Score = 2\frac{RC\ X\ PR}{RC\ +\ PR} \tag{16}$$

Table 3. Comparison of proposed speech recognition technique with existing techniques

Techniques/Performance Metrics	RASTA+LPC+DWT [31]	LPCC+MFCC [32]	SVM [33]	HMM [34]	RNN [35]	Proposed
Accuracy (ACC)	89.5	96.33	87	91	97.49	98.67
Recall (RC)	57.3	96.93	92	93	94.3	98.97
Precision (PR)	83.4	95.73	83	89.5	95.6	98.23
F1-Score	67.93	96.35	87.27	91.22	94.94	98.6

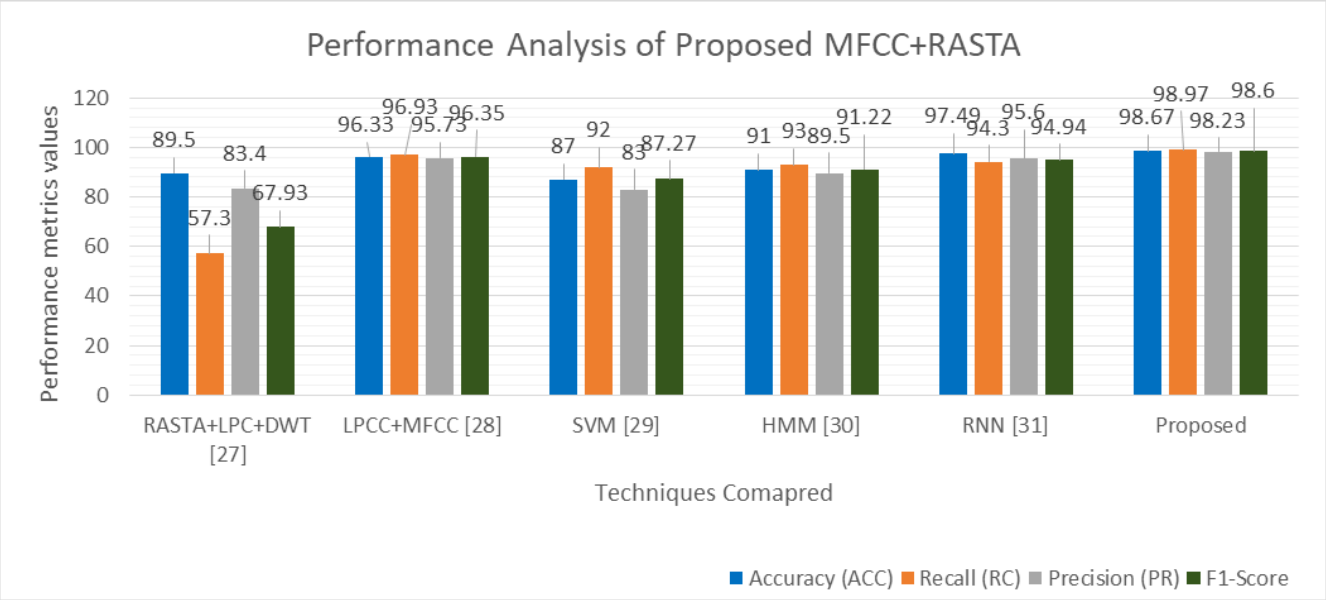

Fig 12. Graphical Analysis of comparison of performance of MFCC+RASTA with existing techniques

The Table 3 records the values of performance parameters calculated for proposed MFCC+RASTA techniques and comparative values of existing techniques. The graphical illustration of comparing proposed technique with existing techniques demonstrated in Figure 12, gives the better place for proposed techniques in terms of accuracy and precision values. Hence, the results demonstrates that the fusion of MFCC and RASTA produces good recognitions in speech detection.

5 Discussion:

The results of the experimentation are tabulated and compared with other existing research works as given in the tables, Tables 1, 2 and 3 . The results are analysed using graphical representations presented in figures, Figure 8 through Figure 12. As mentioned in the explanation of results above, the obtained results are compared with various related researches on speech recognition strategies such as RASTA+LPC+DWT [34], LPCC+MFCC [35], SVM [36], HMM [37] and RNN [38]. The performance evaluation of proposed fusion technique for face recognition is done with various existing techniques proposed by different researchers such as FFT+DWT [27], LBP +DWT +SOM [28], DWT +SVM [29] and RLM for Canny Edge [30]. The results demonstrated that the proposed fusion techniques for both speech and face recognitions placed better in the performance. The proposed model work differently for the data sets. The human face data available inside database and outside database outputs the variations in recognition rates. In the future work, the model will be enhanced to ensure the higher recognition rates for all the possible data sets. Also, the increased number of modalities will be tried for fusion.

6 Conclusion

In this work, both face and voice modalities are discussed along with data sets required for testing face and speech recognition systems. The model is proposed in the paper comprising the face and speech recognition systems, where in face recognition system is implemented by extracting the features of face images using DTCWT and QFT techniques then the fusion of both the techniques is applied. The speech recognition system is implemented by extracting the features of voice data using MFCC and RASTA techniques, and then fusion of both the techniques done to get effective speech recognition system results. The unimodal speech recognition accuracy is 98.67 % and the overall recognition rate of 98.8% is achieved by the proposed model. The results of both the modalities checked and compared with various techniques and demonstrated that the proposed model works better, using performance metrics. In the future work, different biometric traits will be considered, in order to develop system of fusion of more than two biometric modalities, so as to have most advanced and secured human recognition systems.

References

1) Subramanian G, Cholendiran N, Prathyusha K, Balasubramanain N, Aravinth J. Multimodal Emotion Recognition Using Different Fusion Techniques. *2021 Seventh International conference on Bio Signals, Images, and Instrumentation (ICBSII)*. 2021;p. 1–6. doi:10.1109/ICBSII51839.2021.9445146.
2) Saste ST, Jagdale SM. Emotion recognition from speech using MFCC and DWT for security system. *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. 2017;1:701–704. doi:10.1109/ICECA.2017.8203631.
3) Attawibulkul S, Kaewkamnerdpong B, Miyanaga Y. Noisy speech training in MFCC-based speech recognition with noise suppression toward robot assisted autism therapy. *2017 10th Biomedical Engineering International Conference (BMEiCON)*. 2017;p. 1–5. doi:10.1109/BMEiCON.2017.8229135.
4) Assuncao G, Goncalves N, Menezes P. Bio-Inspired Modality Fusion for Active Speaker Detection. *Applied Sciences*. 2021;11:3397. Available from: https://doi.org/10.3390/app11083397.
5) Liu D, Wang Z, Wang L, Chen L. Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning. *Frontiers in Neurorobotics*. 2009;15:8300695–8300695. doi:10.3389/fnbot.2021.697634.
6) Zheng C, Wang C, Jia N. Emotion Recognition Model Based on Multimodal Decision Fusion. *Journal of Physics: Conference Series*. 2021;1873(1):012092–012092. Available from: https://dx.doi.org/10.1088/1742-6596/1873/1/012092.
7) Xie B, Sidulova M, Park CH. Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality the title Fusion. *Sensors*. 2021;21(14):4913–4913. Available from: https://dx.doi.org/10.3390/s21144913.
8) Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*. 2019;7:19143–19165. Available from: https://dx.doi.org/10.1109/access.2019.2896880.
9) Anggraeni D, Sanjaya WSM, Nurasyidiek MYS, Munawwaroh M. The Implementation of Speech Recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machine (SVM) method based on Python to Control Robot Arm. *IOP Conference Series: Materials Science and Engineering*. 2018;288:012042–012042. Available from: https://dx.doi.org/10.1088/1757-899x/288/1/012042.
10) Khusainov AF. Language Models Creation for the Tatar Speech Recognition System. *Indian Journal of Science and Technology*. 2017;10(1). Available from: https://dx.doi.org/10.17485/ijst/2017/v10i1/109954.
11) Adjabi I, Ouahabi A, Benzaoui A, Taleb-Ahmed A. Past, Present, and Future of Face Recognition: A Review. *Electronics*. 2020;10(1):2020–2020. doi:10.3390/electronics9081188.
12) Shanthakumar HC, Nagaraja GS, Basthikodi M. Performance Evolution of Face and Speech Recognition system using DTCWT and MFCC Features. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12(3):3395–3404. Available from: https://dx.doi.org/10.17762/turcomat.v12i3.1603.
13) Maruf MR, Faruque MO, Mahmood S, Nelima NN, Muhtasim MG, Pervez MJA. Effects of Noise on RASTA-PLP and MFCC based Bangla ASR Using CNN. *2020 IEEE Region 10 Symposium (TENSYMP)*. 2020;p. 1564–1567. doi:1109/TENSYMP50017.2020.9231034.
14) Helali W, *Z* Hajaiej, Cherif A. Real Time Speech Recognition based on PWP Thresholding and MFCC using SVM. *Engineering, Technology & Applied Science Research*. 2020;10(5):6204–6208. Available from: https://dx.doi.org/10.48084/etasr.3759.
15) Hidayat R, Bejo A, Sumaryono S, Winursito A. Denoising Speech for MFCC Feature Extraction Using Wavelet Transformation in Speech Recognition System. *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*. 2018;p. 280–284. doi:10.1109/ICITEED.2018.8534807.
16) Tamazin M, Gouda A, Khedr M. Enhanced Automatic Speech Recognition System Based on Enhancing Power-Normalized Cepstral Coefficients. *Applied Sciences*. 2019;9(10):2166–2166. Available from: https://dx.doi.org/10.3390/app9102166.
17) Raju K, Krishna A, Murali M. Automatic Speech Recognition System Using Mfcc-Based Lpc Approach with Back Propagated Artificial Neural Networks. *ICTACT Journal on Soft Computing*. 2020;10(4). doi:10.9790/4200-0606024864.
18) Basthikodi M, Ahmed W. Parallel Algorithm Performance Analysis using OpenMP for Multicore Machines. *International Journal of Advanced Computer Technology (IJACT)*. 2015;4(5):28–32. Available from: https://www.ijact.org/ijactold/volume4issue5/IJ0450005.pdf.
19) Bousnina N, Ghouzali S, Mikram M, Abdul W. DTCWT-DCT watermarking method for multimodal biometric authentication. *Proceedings of the 2nd International Conference on Networking, Information Systems & Security - NISS19*. 2019;19. Available from: https://www.techscience.com/iasc/v27n1/41145/pdf.
20) Shruthi M, Mustafa, Prabhu A. Parellel Implementation of Modified Apriori Algorithm on Multicore Systems. ORALNDO, USA. 2016. Available from: http://www.iiis.org/CDs2016/CD2016Spring/papers/ZA819TX.pdf.
21) Ma Y, Huang Z, Wang X, Huang K. An Overview of Multimodal Biometrics Using the Face and Ear. *Mathematical Problems in Engineering*. 2020;2020:1–17. Available from: https://dx.doi.org/10.1155/2020/6802905.
22) Sarangi PP, Nayak DR, Panda M, Majhi B. A feature-level fusion based improved multimodal biometric recognition system using ear and profile face. *Journal of Ambient Intelligence and Humanized Computing*. 2021. Available from: https://dx.doi.org/10.1007/s12652-021-02952-0.
23) Tomar P, Singh RC. Cascade-based Multimodal Biometric Recognition System with Fingerprint and Face. *Macromolecular Symposia*. 2021;397(1):2000271–2000271. Available from: https://dx.doi.org/10.1002/masy.202000271.

24) Spacek L. Libor Spacek's Facial Images Databases. 2009. Available from: https://cmp.felk.cvut.cz/~spacelib/faces/.
25) Siddiqui MF, Siddique WA, Ahmedh M, Jumani AK. Face Detection and Recognition System for Enhancing Security Measures Using Artificial Intelligence System. *Indian Journal of Science and Technology*. 2020. doi:10.17485/ijst/2020/v13i09/149298.
26) Yale Database. . Available from: http://vision.ucsd.edu/~iskwak/ExtYaleDatabase/ExtYaleB.html.
27) Happy SL, Dasgupta A, George A, Routray A. A video database of human faces under near Infra-Red illumination for human computer interaction applications. *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*. 2012;p. 1–4. doi:10.1109/IHCI.2012.6481868.
28) The ORL Database of Faces. . Available from: http://www.face-rec.org/databases/.
29) Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015;p. 5206–5210. doi:10.1109/ICASSP.2015.7178964.
30) Halvi S, Ramapur N, Raja KB, Prasad S. Fusion Based Face Recognition System using 1D Transform Domains. *Procedia Computer Science*. 2017;115:383–390. Available from: https://dx.doi.org/10.1016/j.procs.2017.09.095. doi:10.1016/j.procs.2017.09.095.
31) Sujatha BM. SOM based Face Recognition using Steganography and DWT Compression Techniques. *International Journal of Computer Science and Information Security*. 2016;14(9):806–826. doi:10.5121/sipij.2016.7304.
32) Sujatha BM, Madiwalar CT, Babu KS, Raja KB, Venugopal KR. Compression Based Face Recognition Using DWT and SVM. *An International Journal (SIPIJ)*. 2016;7(3):45–62. doi:10.5121/sipij.2016.7304.
33) Sujatha BM, Lagali S, Ramapur N, Babu KS, Raja KB, Venugopal KR. Reversible Logic-MUX-Multiplier Based Face Recognition using Hybrid Features. *IOSR Journal of VLSI and Signal Processing*. 2016;6(6):48–64. Available from: http://www.iosrjournals.org/iosr-jvlsi/papers/vol6-issue6/Version-2/F0606024864.pdf.
34) Belahcene M, Laid M, Chouchane A, Ouamane A, Bourennane S. Local descriptors and tensor local preserving projection in face recognition. *2016 6th European Workshop on Visual Information Processing (EUVIP)*. 2016. doi:10.1109/EUVIP.2016.7764608.
35) Maza S, Touahria M. Feature Selection Algorithms in Intrusion Detection System: A Survey. *KSII Transactions on Internet and Information Systems*. 2018;12(10):1–14. doi:10.3837/tiis.2018.10.024.
36) Chen K, Zhou FY, Yuan XF. Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. *Expert Systems with Applications*. 2019;128:140–156. Available from: https://dx.doi.org/10.1016/j.eswa.2019.03.039.
37) Khalvati L, Keshtgary M, Rikhtegar N. Intrusion Detection Based on a Novel Hybrid Learning Approach". *Journal of AI and Data Mining*. 2018;6(1):157–162. doi:10.22044/JADM.2017.979.
38) Acharya N, Singh S. An IWD-based feature selection method for intrusion detection system. *Soft Computing*. 2018;22(13):4407–4416. Available from: https://dx.doi.org/10.1007/s00500-017-2635-2.