CHAPTER 1

INTRODUCTION
1.1 INTRODUCTION

Biometrics is the field of technology devoted to verification or authentication of individuals using biological traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas authentication, a multi-class problem, involves identifying a user from a set of subjects. Hence person authentication is inherently a more difficult task, particularly when the number of registered subjects is large. [1, 82]

In the history of computer science five generation has been passed, each adding a new innovative technology that brings computer nearer and nearer to the people. Now it is sixth generation whose prime objective is to make computer more intelligent, so that it can think like human beings [18, 89]

Recently, there has been a lot of interest in multi-modal biometric person authentication systems [208]. A biometric authentication system verifies the identity of a claimant based on the person’s physical attributes, such as voice, face or fingerprints [55]. Apart from security applications (e.g., access control), authentication systems are also useful in forensic work (where the task is whether a given biometric sample belongs to a given suspect) and law enforcement applications [92, 201]. Speaker authentication is adopted as other technology demands and it can be used with other technologies, such as biometrics or speech recognition. This technique makes it possible to use the person’s features to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers [41, 42].
Broadly speaking, the term information fusion encompasses any area, which deals with utilizing combination of different sources of information, either to generate one representational format, or to, reach a decision. This includes: consensus building, team decision theory, committee machines, integration of multiple sensors, multi-modal data fusion, combination of multiple experts/classifiers, distributed detection and distributed decision making. It is relatively new research area, with pioneering publications tracing back to early 1980’s [132,194,195,254]. Various biometric researches have suggested that no single modality can provide an adequate solution for high security applications. They all point to a common consensus that it is vital to utilize multiple modalities (e.g. visual, infrared, acoustic, chemical sensors, etc.). In order to cope with the limitations of individual biometrics, researchers have proposed using multiple biometric traits concurrently for verification. Such systems are commonly known as multi-modal verification systems. By using multiple biometric traits, systems gain more immunity to intruder attack. For example, it will be more difficult for an impostor to impersonate another person using both audio and visual information simultaneously.

Multi-cue biometrics also helps improve system reliability [239]. For instance, while background noise has a detrimental effect on the performance of voice biometrics, it does not have any influence on face biometrics. On the other hand, while the performance of face recognition systems depends mainly on lighting conditions, lighting does not have any effect on the voice quality. Therefore, audio-visual (AV) biometrics has attracted a great deal of attention in recent years. Another approach to improving the effectiveness of biometric systems is to combine the scores of multiple input samples based on decision fusion techniques [40,83,173]. Although decision fusion is mainly applied to combine the outputs of modality-dependent classifiers, it can also be applied to fuse decisions or scores from a single modality. The idea is to consider
multiple samples extracted from a single modality as independent but coming from the same source. The approach is commonly referred to as multi-sample fusion [51,124].

The need for a reliable identification of interacting users is obvious. At the same time it is well known that the security of such systems is too often violated in everyday life. The possibility to integrate multiple identification cues, such as password, identification card, voice, face, fingerprints and the like will, in principle enhance the security of a system to be used by a selected set of people [177,202].

Speech and speaker authentication is a multimode process, which includes the analysis of uttered acoustic signal and knowledge of grammar semantics and pragmatics. By having image along with audition will improve the recognition score particularly in noisy environment.

1.2 MOTIVATION

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness [164]. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [2,8]. However, today, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. Biometric data is often noisy because of deformable nature of biometric traits, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes. One very
promising approach to improve the overall system's accuracy is to fuse the scores of several biometric systems [98]. Despite many referred works in the literature, it is surprising that there is no coordinated effort in making a benchmark database available for such task [97, 99]. This work is one step towards better sharing of scores to focus on better understanding of the fusion mechanism. In the literature, there are several approaches towards studying fusion. One practice is to use virtual identities whereby a biometric modality from one person is paired with the biometric modality of another person. From the experiment point of view, these biometric modalities belong to the same person. While this practice is somewhat accepted in the literature, it was questioned whether this was a right thing to do or not during the 2003 Workshop on Multimodal User Authentication [123]. The fundamental issue here is the independence assumption that two or more biometric traits of a single person are independent from each other. Another practice is more reasonable: use off-the-shelf biometric systems [140] and quickly acquire scores. While this is definitely a better solution, committing to acquire the systems and to collect the data is admittedly a very time-consuming process. None of the mentioned approaches prevails over the others in understanding the problem of fusion. There are currently on going but independent projects in the biometric community to acquire multimodal biometric databases, e.g., the BANCA [176], XM2VTS [100], BIOMET [205], and MYCT [101] multimodal databases. BANCA and XM2VTS contain face and speech modalities; BIOMET contains face, speech, fingerprint, hand and signature modalities [133]; and MYCT contains ten-fingerprint and signature modalities.

When looking from the point of decision making, there are several motivations for using information fusion [55]:

- Utilizing complementary information (e.g., audio and video) can reduce error rates.
• Use of multiple sensors (i.e., redundancy) can increase reliability.
• Using several cheap sensors rather than one expensive sensor can reduce cost of implementation.
• Sensors can be physically separated, allowing the acquisition of information from different points of view.

Human utilize information fusion everyday [138]. Some examples are: use of both eyes, seeing and touching the same object, or seeing and hearing a person talk (which improves intelligibility in noisy situations) [169]. Several species of snakes combine infrared information with visual information when hunting for prey [139,184].

In literature information fusion is often divided into three main categories, namely, sensor data level fusion (this involves integration of different modalities in raw form e.g., video camera and microphone outputs.), feature level fusion (features are extracted from the raw data and subsequently combined. This involves e.g., audio features for speaker or speech recognition with visual features of the face for face recognition.) and decision fusion (this is the fusion at the most advanced stage of processing and involves combining the decisions of two different classifiers making independent decisions about the identity of the speaker-based on audio and visual features) [50,67]. In general, decision fusion provides a higher degree of robustness, but is accompanied by possible loss of information.

However, it is more intuitive to classify it into two main categories: pre-mapping fusion and post-mapping fusion as shown in Figure 1.2. In pre-mapping fusion, information is combined before any use of classifiers or experts, while in post-mapping fusion; information is combined after mapping from sensor-data/feature space into opinion/decision space. Here, an ensemble of experts or classifiers accomplishes the mapping. While a classifier provides a hard decision, an expert provides an opinion on each possible decision.
Figure 1.2: Non-Exhaustive tree of fusion type.

Silbee and Bovik refer to pre-mapping fusion and post-mapping fusion as pre-categorical integration and post-categorical integration, respectively, while Wark [229] refers to the terms as input level or early fusion and classifier level or late fusion, respectively [169,229].

Fusing the scores of several biometric systems is a very promising approach to improve the overall system's accuracy. Despite many works in the literature, it is surprising that there is no coordinated effort in making a benchmark database available [164]. It should be noted that fusion in this context consists not only of multimodal fusion, but also intramodal fusion, i.e., fusing systems using the same biometric modality but different features, or same features but using different classifiers.

1.3 PREVIOUS WORK IN SPEECH PROCESSING

models for robust speech recognition. R.P.Lippmann [190] gave the comparison between recognition of speech by machines and humans. M.J.F.Gales et. al. [147], used parallel model combination for robust speech recognition. N.Virag [161] performed speech enhancement based on masking properties of the human auditory system. D.Chazen et. al. [62], performed speech reconstruction from mel frequency cepstral coefficients and pitch. Y.J.Chen [257], worked on Conversational Speech Recognition and Verification in Computer Telephony Integration. Tomi Kinnunen et.al. [240] performed Real Time Speaker Identification and Verification. R.Auckenthaler et. al. and B.Xiang et.al., performed text-independent speaker verification with structural gaussian mixture models [38,178]. T.Kinnunen et.al., [54] used pruning algorithm for real time speaker identification. F.Jelinek used Statistical Methods for Speech Recognition [81].

A large number of methods have been proposed for speeding up the verification process. Specifically, Gaussian Mixture Model (GMM) based verification systems [58], [61] have received much attention, since they are considered as the state-of-art method for text-dependent recognition. An efficient GMM-based speaker identification system has also been presented by Pellom and Hansen [30].

The potential applications of speaker identification [33,52] can be found in multi-user system. For instance, in speaker tracking the task is to locate the segments of a given speaker(s) in an audio stream [12], [68], [134], [218].

This particular application concept belongs to the more general group of speaker adaptation methods that are already employed in speech recognition systems [59,186, 252]. Speaker -specific codes in personal speech coding have been also demonstrated to give smaller bit rates as opposed to a universal speaker-independent code [246].
Speaker identification has also been applied to the verification problem [5,65], where the simple rank-based verification method was proposed. For the unknown speaker's voice sample, K nearest speakers are searched from the database. If the claimed speaker is among the K best speakers, the speaker is accepted and otherwise rejected. Similar verification strategy is also used by Beigi et.al., [90].

Speaker identification [110] and adaptation have potentially more application than verification, which is mostly limited to security systems [114]. However the verification problem is still much more studied, which might be due to (1.) lack of applications concepts for the identification problem, (2.) increase in the expected error with growing population size [211], and (3.) very high computational cost.

1.4 PREVIOUS WORK IN IMAGE PROCESSING

Brand and Mason [108] in their paper addressed an issue concerning the current classification of biometrics into either physiological or behavioral. Herpers, Michaelis, Lichtenauer and Sommer developed a method for automatic detection of facial features and characterize anatomical key points [185]. This approach is based on selective search and sequential tracking of characteristic edge and line structures of the facial object to be searched. Toolle, Phillips, Cheng Ross and Wild in their paper evaluated the adequacy of computational algorithms as models of human face processing by looking at how the algorithms and human process the individual faces [19]. Yow and Cipolla detected faces under different scale, orientation and viewpoint [127]. Bronstein et. al., illustrated expression invariant face recognition in three dimension[203]. Palanivel, Venkatesh and Yegnanarayana proposed a method, which used autoassociative neural network models for pattern recognition in speech and image [207]. Verma et. al., used integration of audio visual information for recognition [25]. Shakhnarovich, A.Viota, Moghaddam in
their paper presented progress towards an integrated robust, real-time face detection and demographic analysis system [88]. Li. Gong and Liddel their paper presented a multi-view face model design to extract the shape and pose free texture patterns of face [261]. Bazin and Nixon illustrated probabilistic techniques for face verification [10]. Yang et. al., in their paper illustrated a survey done on face detection [153]. Yow and Cipolla proposed a face detection framework that groups image features into meaningful entities using perceptual organization and assigns probabilities using Bayesian reasoning techniques [128]. Wu & Zhou in their paper discovered eye-analogue segments at a given scale by finding regions, which are roughly as large as real eyes and are darker than their neighborhoods [122]. Turaga and Chew introduced an efficient statistical modeling technique called Mixture of Principal Components (MPC). This model is a linear extension to the traditional Principal Component Analysis (PCA) and uses a mixture of eigen spaces to capture data variations.

1.5 PREVIOUS WORK IN AUDIO-VISUAL PERSON RECOGNITION

The approach in audio-visual person recognition is split into two areas; Non-adaptive and Adaptive approaches [55].

1.5.1. Non-adaptive approaches: -

Fusion of audio and visual information has been applied to automatic person recognition in pioneering papers by Chibelushi et al. in 1993 and Brunelli et al. in 1995 [43,180,181].

Chibelushi et al. combined information from still face profile images and speech using a form of weighted summation fusion:

\[ f = w_1 o_1 + w_2 o_2 \]

where, \( o_1 \) and \( o_2 \) are the opinions from the speech and face profile experts, respectively, with corresponding weights \( w_1 \) and \( w_2 \). Each
opinion reflects the likelihood that a given claimant is a true claimant (i.e., a low opinion suggests that the claimant is an imposter, while a high opinion suggests that the claimant is the true claimant).

Brunelli et al. combined the opinions from a face expert (which utilized geometric features obtained from static frontal face images) and a speech expert using a weighted product approach: [180,181]

\[ f = (o_1)^{w_1} \times (o_2)^{(1-w_1)} \]

When the speech expert was used alone (i.e., \( w_1 = 1 \)), an identification rate of 51% was obtained, while when the face expert was used alone (i.e., \( w_1 = 0 \)), an identification rate of 92% was achieved. Using an optimal weight, the identification rate increased to 95%.

Brunelli et al. used two speech experts (for static and delta features) and three face experts (for the eye, nose and mouth areas of the face) for person identification [180]. The weighted product approach was used to fuse the opinions, with the weights found automatically via a heuristic approach. The static and dynamic feature experts obtained an identification rate of 77% and 71% respectively. Combining the two speech experts increased the identification rate to 88%. The eye, nose and mouth experts obtained an identification rate to 91%. When all five experts were used, the identification increased to 98%.

Dieckmann et al. used three experts (frontal face expert, dynamic lip image expert and text-dependent speech expert) [242]. A hybrid fusion scheme involving majority voting and opinion fusion was utilized. Two of the experts had to agree on the decision and the combined opinion had to exceed a pre-set threshold. The hybrid fusion scheme provided better performance than using the underlying experts alone.

Kittler et al. used one frontal face expert, which provided one opinion for one face image [98]. Multiple images of one person were used to generate multiple opinions, which were then fused by various means, including averaging (a special case of weighted summation fusion). It
was shown that error rates were used by up to 40% and that performance gains tended to saturate after using five images. No results were provided for using more than six images. The results suggest that using a video sequence of the face, rather than one image, provides superior performance.

Kittler et al. attempted to provide theoretical foundations for common fusion approaches such as the summation and product methods [111]. Experimental results for combining the opinions from three experts (two face experts (frontal and profile) and a text-dependent speech experts) showed that the summation approach outperformed the product approach.

Luettin investigated the combination of speech and lip (visual) information using feature vector concatenation [112].

Jourlin et al. used a form of weighted summation fusion to combine the opinions of two experts; a text-dependent speech expert and a text-dependent lip expert. Using an optimal weight, fusion led to better performance than using the underlying experts alone [167,168].

Ben-Yacoub et al. investigated the use of several binary classifiers for opinion fusion using a post-classifier [208]. The investigated classifiers were: Support Vector Machine (SVM), Bayesian classifier (using Beta distributions), Fisher's Linear Discriminant, Decision Tree and Multi Layer Perceptron (MLP). Three experts were used: a frontal face expert and two speech based experts (text-dependent and text-independent). It was found that the SVM classifier (using a polynomial kernel) and the Bayesian classifier provided the best results.

1.5.2 Adaptive approaches: -

Wark et al. extended the work presented in [169,229] by proposing a heuristic method to adjust the weights [235,236]. Experimental results showed that although the performance significantly decreased as the
noise level is increased, it was always better than using the speech expert alone. However, in high noise levels, equal weights (non-adaptive) were shown to provide better performance [171]. A major disadvantage of this method is the calculation of the weights involved finding the opinion of the speech expert for all possible claims (i.e., for all persons enrolled in the system), thus limiting the approach to systems with a small number of clients due to practical considerations (i.e., time taken to verify a claim). Sanderson and Paliwal proposed a weight adjustment method [50].

Hsu, Jain et.al., illustrated the detection using color images [149]. Pathangay and Yenarayana proposed a technique for text-dependent audio-visual person authentication using Dynamic Time Warping (DTW) [243]. Hazen, Weinstein and Park presented the approach for combining face and speaker identification technologies and experimentally demonstrated a fused multi-biometric system, which achieved a 50% reduction in equal error rate over the better of the two independent systems [238]. Fox, Gross, Chazai, F.Cohn and B.Reilly presented a multi-expert system based on the integration of three separate systems employing audio features, static face images and lip motion features [162].

Chen.T, Scalon.P and Reilly.R in their paper showed that the audio expert when fused with visual expert improved the robustness of speech recognition to audio noise [51, 224]. Brunelli. R and Falavigna.D in their multi-expert person identification experiments showed that the integration of visual experts, incorporating visual features from the eyes, nose and mouth, with the audio expert, significantly improved the scores [40,82]. Fox.N, Reilly,R.B and Lucey.S found person identification to be more robust to audio noise when the audio and lip motion experts were integrated.[83,137]
1.6 DIGITAL IMAGE PROCESSING

An image may be defined as a two-dimensional function, \( f(x,y) \), where \( x \) and \( y \) are spatial (plane) coordinates, and the amplitude of \( f \) at any pair of coordinates \((x,y)\) is called the intensity or gray level of the image at that point [199]. When \((x,y)\) and the amplitude values of \( f \) are all finite discrete quantities, we call the image a digital image. The fields of digital image processing refers to processing digital images by means of a digital computer. The digital image is composed of a finite number of elements, each of which has a particular location and value. These elements are referred to as picture elements, image elements, pels, and pixels [199].

1.6.1 Fields of Digital Image Processing

The diagram shown in Figure 1.6.1 does not imply that every process is applied to an image but the intention is to convey an idea of all the methodologies that can be applied to images for different purposes and possibly with different objectives.

- **Image acquisition** is the first process as shown in Figure 1.6.1, where image is in digital form and pre-processing such as scaling is to be done.

- **Image enhancement** is among the simplest and most appealing areas of digital image processing. In this technique the aim is to bring out detail that is obscured, or simply to highlight certain features of interest in an image.

- **Image restoration** is an area that also deals with improving the appearance of an image. However unlike enhancement, which is subjective, image restoration is objective, in the sense that restoration techniques tend to be based on mathematical or probabilistic models of image degradation. Enhancement on the
other hand is based on human subjective preferences regarding what constitutes a "good" enhancement result. [199]

- **Color image processing** is an area that has been gaining importance because of the significant increase in the use of digital images over the Internet. Wavelets are the foundation for representing images in various degrees of resolution.

- **Compression**, as the name implies, deals with techniques for reducing the storage required to save an image, or the bandwidth required to transmit it. Although storage technology has improved significantly over the past decade, the same cannot be said for transmission capacity. This is true particularly in uses of the Internet, which are characterized by significant pictorial content. Image compression is familiar (perhaps inadvertently) to most users of computers in the form of image file extensions, such as the jpg file extension used in the JPEG (Joint photographic experts group) image compression standard.

- **Morphological processing** deals with tools for extracting image components that are useful in the representation and description of shape.

- **Segmentation** procedures partition an image into its constituent parts or objects. In general, autonomous segmentation is one of the most difficult tasks in digital image processing. A rugged segmentation procedure brings the process a long way towards successful solution of imaging problems that require objects to be identified individually. On the other hand, weak or erratic segmentation algorithms almost always guarantee eventual failure. In general, the more accurate the segmentation, the more likely recognition is to succeed. [199]
Figure 1.6.1: Fields of digital image processing.

- *Representation and description* almost always follow the output of a segmentation stage, which usually is raw pixel data, constituting either the boundary of a region (i.e., the set of pixels separating one image region from another) or all the points in the region itself. In either case, converting the data to a form suitable for computer processing is necessary. The first decision that must be made is whether the data should be represented as a boundary or as a complete region. Boundary representation is appropriate when the focus is on external shape characteristics, such as corners and inflections. Regional representation is appropriate when the focus is on internal properties, such as texture or skeletal shape. In some applications, these representations complement each other. Choosing a representation is only part of the solution for
transforming raw data into a form suitable for subsequent computer processing. A method must also be specified for describing the data so that features of interest are highlighted. Description, also called feature selection, deals with extracting attributes that result in some quantitative information of interest or are basis for differentiating one class of objects from another.

- **Recognition** is the process that assigns a label (e.g., "vehicle") to an object based on its descriptors.[199]

### 1.7 FACE RECOGNITION

The subject of face recognition is as old as computer vision due to practical importance of the topic and theoretical interest of cognitive scientists; despite the fact that other methods of identification (such as fingerprints or iris scan) can be more accurate, face recognition always remains a major focus of research due to its non-invasive nature and peoples primary method of personal identification.

Face recognition is a promising means for machine interface and security. For practical use of face recognition, it is necessary to overcome the variation of the face, such as facial expression, face direction, lighting condition and ageing [202]. Face is the most important biometric which plays a major role in conveying identity and emotions. Although the ability to infer intelligence or character from facial appearance is suspicious, the human ability to recognize faces is remarkable. We can recognize thousands of faces learned throughout our lifetime and identify familiar faces at a glance even after years of separation. This skill is quite robust, despite large changes in the visual stimulus due to viewing conditions, expression, aging, and distractions such as glasses, beards or changes in hairstyle. Face recognition does not require active participation of the subject, hence it becomes very attractive biometric. There are many applications of face recognition, which have motivated the interest in this
field. Some of these applications include airport security, access control, criminal identification, authentication of secure systems such as ATM, etc. That is why Automatic Face Recognition has received significant attention from scientists from varying communities of Computer Vision, Image Processing, Neural Network, and Signal Processing.

1.7.1 Techniques of Face recognition

There are different types of face recognition techniques. Some of them are given below:

- Face recognition using eigen images [14].
- Fisher light fields for face recognition across pose and illumination.
- Face recognition using neural networks [16].

In the face recognition using eigen images the facial image of a person at different orientations can be identified. In this method the images, which are in a large database, can be represented in a huge matrix. The mean of the entire face image is then found out by taking average of all images in the database. This mean image is then subtracted from all the images in the database. Then the covariance of eigen values, eigen vectors, and the projections of the images are calculated respectively. Now, each image is a vector of same size as that of the images in the database. The similarity is simply defined as the distance between projections. If this distance is small we say that the images are similar and we can decide which is most similar image in the database.

In many face recognition tasks the pose and illumination conditions of the probe and gallery images are different. In other cases multiple gallery or probe images may be available, each captured from a different pose and under a different illumination. This face recognition algorithm can use any number of gallery images per subject captured at arbitrary poses and under arbitrary illumination. The algorithm operates by estimating the
Fisher light field of the subject's head from the input gallery or probe images. Matching between the probe and gallery is then performed using the Fisher light-fields. This simple linear algorithm can be applied to any number of images, captured from any poses under any illumination. Finally, comparing the probe and gallery, Fisher light fields using a nearest-neighbor algorithm performs matching.
In the third method the image features are extracted using image processing and these features are given to the input of computer simulated ANN model to recognize the image.

1.8 SPEECH RECOGNITION
Speech analysis, synthesis and recognition applications include telephone systems, coding, data compression, voice mail, workstations, personal computers and networks [64]. Speech and audio coding include statistical models, quantization and companding. Speech synthesis includes several forms, such as Formant, articulatory, linear prediction, miscellaneous synthesizers and text to speech systems [225].
One primary objective of speech analysis is to determine the fundamental frequency of voicing, which is often called the pitch [228]. The fundamental frequency of voicing is a physical characteristic that is measured as one would measure the frequency of a sinusoid or a tone [64]: Pitch is a critical feature in speech synthesis and recognition. A plot of the pitch frequency (fundamental frequency of voicing) versus time is often called the pitch contour.
Another objective of speech analysis is to measure the resonances of the vocal tract. These resonances are called the Formant frequencies or more simply the Formants. A plot of Formants versus time is called the Formant tracks or Formant contours. In a similar manner there are gain contours for the gain of the models of the vocal tract as well as contours
for voiced/unvoiced/mixed excitation sounds, nasalization, frication, aspiration and silence. [84]

There are three reasons why we talk. Firstly, in order to express emotion. The pitch, quality, volume, time duration, emphasis etc., express our feelings. Secondly, is to adjust with other people and finally to express our ideas.

Automatic speech recognition and the speaker authentication has been active research area from several decades but the performance achieved by the computers is far from performance achieved by humans. The speech is a complex signal occurring as a result of several transformations at different levels: semantic, linguistic, articulatory and acoustics. Differences in these transformations appear as differences in the spectral properties of the speech signal. Speaker related differences are a result of combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these differences can be used to discriminate between speakers.

The ability of the human listeners to identify the speakers from their voices has been known [148]. On the fundamental level the voices of the two persons differ due to the physical differences between their vocal organs and the manner in which they use them during speech production. For different speakers the anatomical features of the vocal tract are different.

Speech information is primarily conveyed by the short-term spectra, the spectral information contained in an interval of 10-30 ms. The short-term spectra do not completely characterize the speech production process, the information carried by it is basic to many speech-processing systems, including speech and speaker recognition. There are so many methods to characterize short-term spectra, but the dominant features used in previous and current systems are cepstral analysis [9]. Prosodic features,
such as pitch and duration, have been proposed in the past and also methods based on nonlinear discriminate analysis (NLDA) have been evaluated.

There are three main tasks of personnel authentication viz. "Processing", means the enhancing of a human feature." Training" is the most investigated task in the field. The last task arises when it is necessary to determine name and other information about a person just based on his/her one causal feature [211]. Due to differences between the tasks there is not a universal approach or algorithm for personnel authentication.

The past seven years have seen no great change in the feature selection component of speaker recognition systems. As a matter of fact, the methods of VQ (Vector Quantization), DTW (Dynamic Time Warping), and NN (Nearest Neighbors) are now less common than HMM (Hidden Markov Models) and ANN (Artificial Neural Networks).

Several independent techniques on personnel authentication from their images and sound features are available in the literature. Most of these techniques discriminate the person from the pattern analysis in sound data [60]. Many of the techniques presently used are dependent on the picture movement in video and analysis of eigen values. These techniques are not generic in its true sense.

1.9 FEATURE EXTRACTION

The speech information for speaker authentication should use the same language and a common code from a common set of speech sounds. The ideal speech characteristic [251] for the speaker identification should posses the following properties:

1. They should occur naturally and frequently in the normal speech.

2. Take less space for storage of data.
3. Use of faster or less complex classification procedures.
4. Achievement of lower error.
5. Easily measurable.
6. It should not change over with time.
7. The speech features should vary as much as possible among the speakers but be as consistent as possible for the particular speaker.
8. It should not be affected by the background noise.

The speech feature extraction using DSP toolbox in MATLAB consists of transforming the raw data into concise representation that contains the relevant information and is robust to the variations. The following speech features are extracted and analyzed:

1. Formant Frequency F1 to F4.
3. Time Duration.
4. Number of Zero Crossing.
5. Pitch & Pitch Amplitude.
6. Maximum Amplitude i.e. Ceps Max.

The image features are extracted by Image Processing toolbox in MATLAB. The following features of image are extracted and analyzed:

1. Length of Eye1.
2. Width of Eye1.
3. Center Dimension of Eye 1.
4. Center Dimension of Eye 2.
5. Length of Eye2.
7. Length of Mouth.
8. Width of Mouth.
9. Center Dimension of Mouth.
10. Distance Between Center Points of Eye 1 & Eye 2.
11. Distance Between Center Points of Eyes & Mouth.

1.10 ARTIFICIAL NEURAL NETWORK (ANN)
An Artificial Neural Network (ANN) is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example [74]. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true for ANNs as well.

Artificial Neural networks have emerged as a powerful tool for person authentication technique [255]. The need for this technique arises whenever computer interacts with the real world. The Neural Network can realize the spatial, temporal or any other relationship that can perform the task such as classification, prediction and function estimation. The Artificial Neural Networks (ANN) provides the potential of an alternative information-processing paradigm that involves large inter-connected networks of processing units (PE). It resembles the brain processor in two respects:

1. The network through the learning process acquires knowledge.
2. Interconnecting strengths known as synaptic weights are used to store the knowledge.

ANNs were inspired by models of biological neural networks since much of the motivation came from the desire to produce artificial systems capable of sophisticated, perhaps "intelligent", computations similar to
those that the human brain routinely performs, and there by possibly to enhance our understanding of the human brain [142]. The network learns the similarities among patterns directly from instances of them. That is, they infer solution from the data without prior knowledge of the regularities in the data; they extract the regularities empirically.

The procedure starts with selecting, analyzing and manipulating data, often using techniques from statistics and signal processing [250]. The Neural Networks has the advantage that they can infer subtle, unknown relationship from data; this characteristic is useful because gathering of the data does not require explaining. The network is generalized and this is useful, because the real-world data is noisy, distorted and often incomplete. The network is non-linear, that is, they can solve some complex problems more accurately than linear techniques can do. The highly parallel characteristic of the Neural Network [157] helps to perform many identical, independent operations that can be executed simultaneously because parallel hardware can make neural network faster than any other alternative method. Several features distinguish this paradigm from conventional computing and traditional artificial intelligence approach, which are as follows:

1. In ANN information processing is inherently parallel.
2. Knowledge is distributed throughout the system.
3. Extremely fault tolerant.
4. Adaptive model free function estimation and non-algorithmic strategy.

ANNs find applications in multi-disciplinary field including: Finance; Industry; Agriculture; Business; Physics; Statistics; Cognitive sciences; Neuro-sciences; Weather forecasting; Computer science and Engineering; Spatial analysis and Geography.
The neural network has to be structured for a particular application, and the network has to be trained. To start this process the initial weights are chosen randomly. Then, the training, or learning, begins [215]. There are two approaches to training - supervised and unsupervised. Supervised training [200] involves a mechanism of providing the network with the desired output either by manually "grading" the network's performance or by providing the desired outputs with the inputs. Unsupervised training is where the network has to make sense of the inputs without outside help. The vast bulk of networks utilize supervised training. Unsupervised training is used to perform some initial characterization on inputs. [237].

1.11 OVERVIEW OF PRESENT WORK
In speech processing the spectrogram has been plotted and the Formant frequencies (F1 to F4) are extracted. The Power Spectral Density curves are plotted from which peak and average value of the spectral density has been calculated. The Cepstral analysis gives the Ceps Max; pitch and Number of zero crossing features of the speech has been extracted with the help of the MATLAB software.
The feed forward neural network up to 3-layers is created and simulated with the help of neural network toolbox on MATLAB 7.0.1. The well-known Back Propagation training algorithm is used to train the network.
The acoustic speech signals [120] are susceptible to the noise as all real world applications are subject to some form of noise. Speaker authentication by machine using analysis of the face and voice are the two most dominant modalities. These two has been treated separately for visual face recognition [166] and the acoustic speaker recognition.
In the face recognition using image processing the facial image of a person can be identified. The images, which are in a large database, can be represented by huge matrix. The sequence of process is carried out to
acquire images, i.e., convert Original image into Gray image, Binary gradient mask at previous image, Dilated gradient mask, Filled holes, Cleared border in image, Segmented image, Boundary extraction of image, which take out line of the images. Extracting features like position and size of eyes and mouth and distance between them are calculated.

In the present work (model shown in Figure 1.11) the fusion of the features extracted from the images and speech utterances of the speakers belonging to the same age groups is made. These features are given to the computer simulated ANN model, which consists of feed forward perceptron model and is trained by well-known Back Propagation Algorithm (BPA). The ANN model output authenticate the speaker identities in least number of training cycles and minimum input features of speech and image. This method is robust and do not get affected even if image is distorted and speech signal have noise. By fusion technique the distorted images and noisy signals of sound with less number of image and speech features given to input model of ANN gives better than 98% recognition score.
Figure 1.11: Model for Person authentication