Chapter 8

Summary and Conclusions

8.1 Summary of the work

In this thesis, methods were proposed for person authentication using speech, face and visual speech modalities. The performance of the system was evaluated for newsreaders in TV broadcast news data. During news reading, the background around the newsreader is almost constant accompanied by a small motion of the reader in the foreground. Hence, the motion information was used to estimate the face region.

The eye region has low luminance, low red chrominance and high blue chrominance when compared to the forehead region of the face. Using these facts the face region was processed to determine the locations of the eyes. The facial features were extracted relative to the locations of the eyes for each frame in the video. The multiscale morphological erosion operation was used for facial feature extraction. The distribution of the facial features of a subject was captured using an AANN model. For testing the identity claim of a subject, each facial feature vector is given as input to the claimant model. The output of the model is compared with the input to compute the normalized squared error. The error is transformed into a confidence score. If the average confidence score is greater than a threshold then the claim is accepted, otherwise it is rejected.

The color distribution of the nonlip region of the face is different from the lip region. The color distribution of the pixels in the nonlip region was captured using a
Gaussian distribution, and it was used estimate the center of the mouth. The nonlip region was extracted relative to the locations of the eyes. The static nature of the visual speech features was extracted relative to the locations of the eyes and mouth using multiscale morphological dilation operation. The distribution of these features of a subject was captured using an AANN model. Similarly, the temporal nature of the visual speech features were extracted using morphological dilation operation. The temporal information present in the visual features of a subject was learned using HMM.

The acoustic features were derived from the speech signal, and were represented by WLPCC. The distribution of the acoustic features of a subject was captured using an AANN model.

The evidence from speech, face and visual speech (static) modalities were combined at the matching score level using a weighted rule, and the result was used to accept or reject the identity claim of a subject. The performance of the method was compared with the existing audio-video based person identification or authentication methods. The method satisfies the following requirements for a person authentication technique:

- Invariant to size of the face, its position in the image and its background.
- Invariant to orientation and pose of the face to some extent.
- A subject can be enrolled into the system without using the features of other subjects (discriminating information was not used). Similarly, other subject models or scores were not used for authentication.
- Able to authenticate the identity claim of a subject within a reasonable time.
8.2 Major Contributions of the Work

The most important contribution of the research reported in this thesis is that it presents an automatic audio-video based person authentication method using speech, face and visual speech modalities. The major contributions of this thesis are:

- A multimodal biometric system was proposed for person authentication which is more effective and robust than using a unimodal biometric system based on speech or face.

- The distribution capturing ability of the autoassociative neural network models was exploited to capture the distribution of feature vectors describing each of the biometric modalities such as speech, face and visual speech. These distributions are unique for each individual.

- New methods were proposed to derive the feature vectors for face and visual speech. These feature vectors are invariant to size of the face, its position in the image and its background, and orientation of the face.

- A method based on $YC_rC_b$ representation of the face image was proposed for estimating the locations of the eyes. The locations of the eyes were used to derive the size and orientation invariant features for face and for visual speech.

- The temporal nature of the visual speech was analyzed for person authentication using hidden Markov models.

- A new method of tracking the face region from a video sequence was proposed for real time face tracking, and the method is invariant to the size of the face, background and lighting conditions.
8.3 Directions for Future Research

The method proposed in this thesis for automatic person authentication addresses the issues such as size of the face, its position in the image and its background, computational complexity, orientation and pose of the face to some extent. The method is less sensitive to variation in the image brightness and it is sensitive to shadows and other lighting conditions.

The method described in this thesis does not consider the issues like variation in lighting conditions and the effects of noise. The performance of the system degrades if the lighting condition and the acoustic environment during training and testing are different. The performance of the system must be invariant to wide range of lighting and acoustic environments encountered in the real world. Further efforts are required for extracting facial and acoustic features that are invariant to lighting and acoustic environments.