Chapter 9

Conclusion and Future works

9.1 Concluding Remarks - Summary of the core contribution of the thesis

Emotion recognition through acoustic features is useful and advantageous to recognize the emotions of a speech sample. In this research work, we ascertain that prosodic features together with spectral features (feature fusion) increases the performance of the emotion recognition system. But the question of dimensionality of the fused feature vector is very high with feature fusion, due to which the performance of the system could not reach to an optimal state. So we go for optimal feature subset selection methods with which the dimensionality of the fused feature vector is reduced by eliminating the irrelevant and unwanted features. This would increase the performance of the system significantly, with a low dimensional feature vector which contains more important emotion specific information. The feature selection plays a vital role in recognizing the emotion besides increasing the performance of the system. In this task, we emphasize on performing the evaluation over different emotional databases with various classification techniques to check the robustness of the emotion recognition system.

9.2 Summary of the work

In this work, we could solve all the research questions. The following are the answers to the research questions.
1. Which acoustic features are best for emotion classification in speech emotion recognition system?

An extensive survey has been conducted on existing acoustic features. An important task in speech emotion recognition system is in the selection of features which best classify the emotions in speech emotion recognition system. There is no theoretical explanation on which features best classify the emotion among the existing features because research work in this field has been carried out on various features and their combinations with different databases applying various classification techniques. No two works are similar and alike with respect to all the aspects i.e when the databases are similar, classifiers chosen are different or when the classifiers chosen are similar, features selected are different.

More number of acoustic features are extracted from the speech sample, among these some of the features contain more emotion specific information and the others contain very little emotion specific information of the speech sample. i.e the use of all the existing features are not enough for emotion recognition. The answer to this question is the proper selection of subset of existing features which best can classify the emotion of the speech sample. Initially, speech sample is divided into a set of frames and features are computed for each frame and these extracted features are analysed over simple statistical functions like mean, variance, minimum, range, skewness and kurtosis. These statistical values are given as input to the classifier for the recognition of the emotion of the speech sample.

2. what combination of features are beneficial in extracting the emotion of the speech sample?

The acoustic features are divided into prosody and spectral. Some of the
features come under prosody and the other come under spectral, based on vocal folds and vocal tract activities. The selection of features for this work depends on the study of various features along with their pros and cons. In this work we considered the combination of both prosody and spectral features which comes under feature fusion. The prosody and spectral features considered for this work are Energy, Pitch and Mel Frequency Cepstral Coefficients.

The question that arises is whether the features alone are effective or the combination of features are effective in recognizing emotions. i.e in another way the question is the role of prosody and spectral features and the importance of their combinations in identifying the emotional expressions in speech emotion recognition system. The answer to this question is that the implementation of speech emotion recognition system with both individual features as well as the combination of features(feature fusion). The results suggest that the feature fusion yields better results than that of the individual features.

3. which classifiers are most useful in estimating the emotion of the speech sample.

Similar to the selection of features which best classify the emotion, the selection of classifiers which best classify the emotion of the speech sample is also a difficult task. i.e there is no clear agreement on which are the best classifiers for emotion classification. The reason for this is, though the classifiers thus considered are same the features selected and the databases collected or chosen are different.

There are several classifiers in estimating the emotion of the speech sample like Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM),
Artificial neural networks, Support Vector Machine (SVM) and k Nearest Neighbor (kNN) and many others. The question that arises here is the importance of each classifier in detecting the emotion of the speech sample. The answer to this question is the performance of the classifier depends on the dimensionality of feature vector and preprocessing of the speech sample. That is why, in this work we have considered four classification techniques viz., LDA, RDA, SVM and kNN for the identification of the emotion of the speech sample and also an extensive comparative study has been carried out on these classifiers.

4. which algorithms are useful in selecting an optimal set of features. i.e how the dimensionality of the speech sample effect the efficiency of the speech emotion recognition system.

The dimensionality of the speech sample plays a major role in estimating the performance of the speech emotion recognition system. The more the dimensionality of the feature vector, the less the performance of the system because it suffers from ‘small sample size problem’. i.e the number of features are more when compared with the number of speech samples available. Now the solution for this problem is to reduce the dimensionality of the feature vector in such a way that it might not decrease the performance of the system.

There are several algorithms in selecting the optimal subset of features from a high dimensional feature vector. Now our task is to select the feature selection algorithms which are best suited in extracting an optimal feature subset. In this work we have selected Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) which are best suitable in the selection of an optimal feature subset by eliminating the repeated, irrelevant and less emotion specific information from the fused feature vector.
5. what methods are proposed in improving the performance of the speech emotion recognition system.

Initially we have implemented speech emotion recognition system with prosody and spectral features separately. To improve the performance of the speech emotion recognition system further, a feature fusion technique is proposed. The performance of the system is improved with this feature fusion technique. As it will not reach to an optimal state, to improve the performance of the speech emotion recognition system further, a two stage feature selection method is proposed. The selection and fusion of features comes under first stage feature selection.

In the second stage feature selection method, these fused features are given as input to the optimal feature subset selection methods. These methods generate an optimal feature subset, which will improve the performance of the speech emotion recognition system effectively. This optimal feature subset is given as input to the classifier which will classify the emotional state of the speech sample more accurately with which the performance of the speech emotion recognition system reaches to an optimal state.

6. what is the significance of speech emotion recognition system in the context of a driver or how the acoustic features are helpful in recognizing the emotion of the driver dynamically.

In the context of Advanced Driver Assistance System, Pitch, Zero crossing rate, Energy and Mel frequency cepstral coefficients are used. During the course of driving, first we have to classify the noise and the driver’s acoustic information. For this purpose zero crossing rate and energy are used. If the energy is low and zero crossing rate is high, the speech signal becomes noisy or unvoiced one. Based on zero crossing rate and energy, the voice
signals are identified. The pitch and Mel frequency cepstral coefficients are calculated from the acoustic information to recognize the emotion.

In real time driving situation, the entertainment systems and the co-passenger’s voice decrease the accuracy of the speech emotion recognition system. So to improve the performance of the system, we have to filter particularly the driver’s voice from the other sounds. After the extraction of driver’s voice, speech emotion recognition system will be able to identify the emotional state of the driver and make him alert from causing an accident.

9.3 Future works

This work can be extended by adding some more useful features to get more emotion specific information with which the performance of the system can be improved. In a similar way, selection of a genetic algorithm for the optimal feature subset selection can also be used as another addition for the present work. On the other hand, this work can be improved in the direction of reducing the acoustic noise generated by the vehicle as well as the vehicle entertainment systems to improve the quality of the acoustic sensor information.

This work can also be extended with image and video processing along with speech processing by using feature fusion. Among the existing pattern recognition methods like Statistical Approach, Syntactical Approach, Template Matching and Neural Networks. Statistical Approach based pattern recognition techniques are applied in this thesis because this approach deals with high dimensional data very effectively and efficiently. Further, this work can also be extended to the remaining pattern recognition techniques.