# Chapter 6
# Summary and Conclusion

## 6.1 Summary of the Book

In this work, articulatory and excitation source features are explored for improving the performance of phone recognition systems (PRSs). Methods are proposed to extract articulatory and excitation source features from the given speech signal. Pattern recognition models such as hidden Markov models (HMMs) and feedforward neural networks (FFNNs) are explored for deriving the articulatory features (AFs) from the speech signal. The excitation source information present in the linear prediction (LP) residual of the speech signal is captured using two sets of features. It is observed that the use of either AFs or excitation source features along with the spectral features improves the performance of PRSs. The improvement achieved using combination spectral and AFs is much higher compared to the improvement obtained using the combination of spectral and excitation source features. It is found that the excitation source features can be used for improving the robustness of PRSs. In this work, HMMs are used for building PRSs. TIMIT and Bengali speech corpora are used for evaluating the proposed features. The proposed features and models are also evaluated on read, extempore, and conversation modes of speech in Bengali. TIMIT PRSs are developed using 48 phones. The number of phones used for developing PRSs of read, extempore, and conversation modes of speech in Bengali language are 35, 31, and 31, respectively [1, 2]. Mel frequency cepstral coefficients (MFCCs) containing vocal tract information are used as spectral features. The tandem PRSs are developed by using FFNNs in the first stage to derive phone posteriors, and HMMs in the second stage for mapping the combination of spectral and posterior features to phone identities.

The articulatory features are explored for improving the performance of PRSs. Five AF groups, namely (i) place, (ii) manner, (iii) roundness, (iv) frontness, and (v) height, are considered. AFs for each AF group are derived by training separate FFNNs for each AF group. Five different AF-based tandem PRSs are developed using the combination of MFCCs, and AFs derived for each AF group. Hybrid PRSs are



K.S. Rao and Manjunath K.E., *Speech Recognition Using Articulatory and Excitation Source Features*, SpringerBriefs in Speech Technology,
DOI 10.1007/978-3-319-49220-9_6

developed by combining the evidences from AF-based tandem PRSs using weighted combination approach. The performance of hybrid PRSs is compared with the baseline PRS and phone posteriors (PP)-based tandem PRS. Hybrid PRS developed using evidences from all five AF groups is having higher performance compared to the hybrid PRS developed using evidences from subset of five AF groups. It is found that the hybrid PRS developed using AFs from all the five AF groups outperforms the conventional PP-based tandem PRS. PP- and-All-AF-based hybrid PRS has shown highest recognition accuracy. The highest improvement obtained in the recognition accuracy of read, extempore, and conversation modes of speech is 7.13, 6.66, and 6.95%, respectively. TIMIT PRS has shown an improvement of 6.31% in recognition accuracy. Read speech has shown highest improvement in the recognition accuracy. The improvement in performance of extempore and conversation modes of speech are almost same. The AFs are mainly responsible for improving the performance of read and extempore modes of speech, whereas the improvement in the performance of conversation speech is mainly due to PPs [3, 4].

The excitation source information is parameterized using two techniques: residual Mel frequency cepstral coefficients (RMFCCs) and Mel power differences of spectrum in sub-bands (MPDSS). The use of excitation source information in addition to vocal tract information has improved the performance of PRSs in all three modes of speech. The PRSs developed using only excitation source information have lower recognition accuracy compared to the PRSs developed using vocal tract information alone [5]. Among the three Bengali PRSs developed using excitation source features, the extempore speech PRS has shown highest improvement in the performance, while the conversation speech PRS has shown least improvement [4]. The combination of spectral and excitation source features is used for developing robust PRSs. The robustness of the proposed excitation source features in phone recognition is analyzed using white and babble noisy speech samples. It is found that the performance of PRSs is higher in case of additive babble noise than that of additive white noise [6].

## 6.2 Contributions of the Book

The major contributions of this work can be summarized as follows:

- Speech data in read, extempore, and conversation modes of Bengali language is collected and manually transcribed using *international phonetic alphabet* chart.
- Methods are proposed to derive the articulatory features from the spectral features using FFNNs.
- The development of *phone recognition systems* using combination of spectral and articulatory features is proposed.
- Methods are proposed to capture the excitation source information from the LP residual of the speech signal.

- The development of *phone recognition systems* using combination of spectral and excitation source features is proposed.
- The articulatory and excitation source features are analyzed across read, extempore, and conversation modes of speech.

## 6.3 Future Scope of Work

- In this book, articulatory and excitation source features are explored separately to improve the performance of PRSs. In future, the combination of articulatory and excitation source features can be explored to improve the performance of PRSs.
- In this study, AFs are derived from the spectral features using FFNNs. Instead, the AFs derived from signal processing techniques can be explored for improving the performance of PRSs. The signal processing techniques such as modified group delay function, strength of excitation derived from zero-frequency filtered signal can be used to derive AFs.
- In this work, the discriminative features, which are used in developing tandem PRSs, are derived using FFNNs. The discriminative classifiers such as support vector machines (SVM) can be explored instead of FFNNs.
- In this book, articulatory and excitation source features are explored for developing phone-based speech recognition systems. Instead, syllable-based speech recognition systems can be considered to demonstrate the performance improvement using articulatory and excitation source features.
- In this work, we have considered LP residual signal as excitation signal. In future, the *glottal volume velocity* can be considered as excitation signal, and similar study can be carried out.
- In this work, LP residual signal is parameterised using RMFCCs and MPDSS. One can explore other parametrization techniques such as the *glottal flow derivative parameters* to parameterize the LP residual signal.
- In this study, articulatory and excitation source features are used for improving the performance of HMM-based PRSs. In future, the articulatory and excitation source features can be used for improving the performance of PRSs developed using deep neural networks.
- In this book, we have analyzed the robustness of excitation source features using additive white and babble noises with fixed SNR. However, in real-life applications, the test samples may be degraded by various background noises with different SNRs. In future, this work can be extended with varying noise types and noise levels.
- Proposed articulatory and excitation source features may be explored for other Indian languages. The variations in recognition accuracies across different Indian languages can be analyzed.
- By exploiting the availability of transcribed speech in multiple Indian Languages, the performance of individual PRSs (i.e. the PRS of each language) may be improved.

## References

1. Manjunath K.E., K. Sreenivasa Rao, D. Pati, Development of phonetic engine for indian languages: bengali and oriya, in *IEEE International Oriental COCOSDA* (2013)
2. Manjunath K.E., K. Sreenivasa Rao, Automatic phonetic transcription for read, extempore and conversation speech for an indian language: bengali, in *IEEE National Conference on Communications* (2014)
3. Manjunath K.E., K. Sreenivasa Rao, M. Gurunath Reddy, Two-stage phone recognition system using articulatory and spectral features, in *IEEE International Conference on Signal Processing and Communication Engineering Systems* (2015), pp. 107–111
4. Manjunath K.E., K. Sreenivasa Rao, Source and system features for phone recognition. Int. J. Speech Technol. 1–14 (2014)
5. Manjunath K.E., K. Sreenivasa Rao, M. Gurunath Reddy, Improvement of phone recognition accuracy using source and system features, in *IEEE International Conference on Signal Processing and Communication Engineering Systems* (2015), pp. 501–505
6. Manjunath K.E., K. Sreenivasa Rao, Articulatory and excitation source features for speech recognition in read, extempore and conversation modes. Int. J. Speech Technol. 1–14 (2015)