SPRINGER BRIEFS IN
ELECTRICAL AND COMPUTER ENGINEERING

K. Sreenivasa Rao
Manjunath K.E.

Speech Recognition Using Articulatory and Excitation Source Features

Springer

SpringerBriefs in Electrical and Computer Engineering

Speech Technology

Series editor

Amy Neustein, Fort Lee, NJ, USA

Editor's Note

The authors of this series have been hand-selected. They comprise some of the most outstanding scientists—drawn from academia and private industry—whose research is marked by its novelty, applicability, and practicality in providing broad based speech solutions. The SpringerBriefs in Speech Technology series provides the latest findings in speech technology gleaned from comprehensive literature reviews and *empirical investigations* that are performed in both laboratory and *real life* settings. Some of the topics covered in this series include the presentation of real life commercial deployment of spoken dialog systems, contemporary methods of speech parameterization, developments in information security for automated speech, forensic speaker recognition, use of sophisticated speech analytics in call centers, and an exploration of new methods of soft computing for improving human-computer interaction. Those in academia, the private sector, the self service industry, law enforcement, and government intelligence, are among the principal audience for this series, which is designed to serve as an important and essential reference guide for speech developers, system designers, speech engineers, linguists and others. In particular, a major audience of readers will consist of researchers and technical experts in the automated call center industry where speech processing is a key component to the functioning of customer care contact centers.

Amy Neustein, Ph.D., serves as Editor-in-Chief of the International Journal of Speech Technology (Springer). She edited the recently published book "Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics" (Springer 2010), and serves as quest columnist on speech processing for Womensenews. Dr. Neustein is Founder and CEO of Linguistic Technology Systems, a NJ-based think tank for intelligent design of advanced natural language based emotion-detection software to improve human response in monitoring recorded conversations of terror suspects and helpline calls. Dr. Neustein's work appears in the peer review literature and in industry and mass media publications. Her academic books, which cover a range of political, social and legal topics, have been cited in the Chronicles of Higher Education, and have won her a pro Humanitate Literary Award. She serves on the visiting faculty of the National Judicial College and as a plenary speaker at conferences in artificial intelligence and computing. Dr. Neustein is a member of MIR (machine intelligence research) Labs, which does advanced work in computer technology to assist underdeveloped countries in improving their ability to cope with famine, disease/illness, and political and social affliction. She is a founding member of the New York City Speech Processing Consortium, a newly formed group of NY-based companies, publishing houses, and researchers dedicated to advancing speech technology research and development.

More information about this series at http://www.springer.com/series/10043

K. Sreenivasa Rao · Manjunath K.E.

Speech Recognition Using Articulatory and Excitation Source Features

Springer

K. Sreenivasa Rao
Department of Computer Science
and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, West Bengal
India

Manjunath K.E.
Indian Institute of Technology Kharagpur
Bangalore, Karnataka
India

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Electrical and Computer Engineering
ISSN 2191-737X ISSN 2191-7388 (electronic)
SpringerBriefs in Speech Technology
ISBN 978-3-319-49219-3 ISBN 978-3-319-49220-9 (eBook)
DOI 10.1007/978-3-319-49220-9

Library of Congress Control Number: 2016958474

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The goal of developing a phone recognition system (PRS) is to derive the sequence of basic sound units from the speech signal. Most of the state-of-the-art PRSs are developed using spectral features such as Mel frequency cepstral coefficients. Spectral features mainly represent the gross shape of the vocal tract, but not the information related to the excitation source or the positioning and movements of various articulators. But, the production of each sound unit is characterized by articulatory and excitation source features in addition to vocal tract features. It is impossible to produce a sound unit without having an appropriate source of excitation. The rate of vibration of vocal folds varies from one phone to another phone based on their inherent characteristics as well as the influence of coarticulation characteristics due to the presence of adjacent phones. The positioning and movement of various articulators during the production of a sound unit change from one sound unit to another. A unique combination of articulators in the vocal tract and specific source of excitation results in production of a particular sound unit.

In this work, the articulatory and excitation source features are explored for improving the performance of PRSs. The articulatory features (AFs) are derived from the spectral features using feedforward neural networks (FFNNs). Five AF groups, namely manner, place, roundness, frontness, and height, are considered. Five different AF-based tandem PRSs are developed using the combination of spectral features and AFs derived from FFNNs of each AF group. The systematic analysis of phone-level accuracies contributed by each AF group is carried out. Hybrid PRSs are developed by combining the evidences from AF-based tandem PRSs using weighted combination approach. It is observed that the use of AFs in addition to spectral features has lead to improvement in the performance of PRSs.

The excitation source information is derived by processing linear prediction (LP) residual of the speech signal. The use of excitation source information has shown improvement in the performance of PRSs. The robustness of proposed excitation source features is demonstrated using white and babble noisy speech samples. The PRSs developed using the combination of vocal tract and excitation source features are more robust to noise than the PRSs developed using vocal tract features alone. The performance of tandem PRSs is improved using excitation

source features in addition to spectral features. The performance of PRSs developed using articulatory and excitation source features across read, extempore, and conversation modes of speech is analyzed, and results are compared. The use of articulatory and excitation source features has shown improvement in all the three modes of speech.

This book is mainly intended for researchers working on speech recognition area. This book is also useful for the young researchers, who want to pursue research in speech processing with an emphasis on articulatory and excitation source features. Hence, this may be recommended as the text or reference book for the postgraduate level advanced speech processing course. This book has been organized as follows:

Chapter 1 introduces basic concepts of speech recognition and its applications. The articulatory and excitation source features are described briefly. Chapter 2 provides compendious reviews about the use of articulatory and excitation source features to develop speech recognition systems. Chapter 3 discusses the proposed approaches to derive and use AFs for phone recognition task. The development of tandem and hybrid PRSs using AFs is proposed. Chapter 4 describes the proposed methods to parameterize and use the excitation source information to perform phone recognition. The use of excitation source features to improve robustness of PRSs is also discussed. Chapter 5 investigates the use of articulatory and excitation source features to improve performance of PRSs across read, extempore, and conversation modes of speech. Chapter 6 provides a brief summary and conclusion of this book with a glimpse toward the scope for possible future work.

We would especially like to thank all professors of computer science and engineering, IIT Kharagpur for their moral encouragement and technical discussions during course of editing and organization of this book. Special thanks to our colleagues at Indian Institute of Technology Kharagpur, India, for their cooperation to carry out the work. We are grateful to our parents and family members for their constant support and encouragement. Finally, we thank all our friends and well-wishers.

Kharagpur, India K. Sreenivasa Rao
Bangalore, India Manjunath K.E.

Contents

Acronyms

AF	Articulatory feature
ASR	Automatic speech recognition
CA	Classification accuracy
DCT	Discrete cosine transform
DFT	Discrete fourier transform
FFNN	Feedforward neural network
GMM	Gaussian mixture models
HMM	Hidden Markov model
Hz	Hertz
IFT	Inverse fourier transform
IPA	International Phonetic Alphabet
LP	Linear prediction
LPCC	Linear prediction cepstral coefficients
LPR	Linear prediction residual
MFCC	Mel frequency cepstral coefficients
MPDSS	Mel power difference of spectrum in sub-bands
ms	Milli seconds
PDSS	Power differences of spectrum in sub-bands
PP	Phone posterior
PPRT	Phonetic and Prosodically Rich Transcribed
PRS	Phone recognition system
RMFCC	Residual Mel frequency cepstral coefficient
SNR	Signal-to-noise ratio
SVM	Support vector machines
TIMIT	Texas Instruments and Massachusetts Institute of Technology

Chapter 1
Introduction

1.1 Speech Recognition Systems

In the present world, it is very hard to live without any interaction with machines. Major way of communication between the man and the machine is happening through hardware interfaces such as keyboard, mouse, and touch screen. The most convenient way of communication for humans is speech. Hence, it would be very efficient to have speech mode of communication between the man and the machine. This can be accomplished by developing speech recognition (speech understanding) [1–3] and speech synthesis (speech generation) systems [4–6]. Speech recognition systems are used for decoding the message conveyed in the speech signal into text. Speech recognition systems can be either phone-based or syllable-based. Phones represent the basic sound units in a language, whereas the syllables represent the sound units composed of a central nucleus, which is mostly a vowel, with optional initial and final consonants. In this book, we have discussed the phone-based speech recognition systems called phone recognition systems (PRSs). The purpose of developing a PRS is to derive a sequence of basic sound units from the speech signal. PRS has a wide range of applications in domains such as health care, military, telephony, dictation, robotics, and home automation. The PRS is used in developing systems for speech-to-text conversion, language recognition, and audio search engine.

The machine learning approaches such as hidden Markov models (HMMs), feed-forward neural networks (FFNNs), and support vector machines (SVMs) are used to develop PRSs. HMMs are used to model the sequence of vocal tract shapes contributed to the production of sound unit, with local spectral variability modeled using mixtures of Gaussian densities [7]. The FFNNs and SVMs have good discriminative power to distinguish between correct output class and the rival ones [8]. In the context of phone recognition, discrimination between vocal tract shapes offered by various sound units is exploited by the FFNNs and SVMs. Generally, the standard spectral features such as linear prediction cepstral coefficients (LPCCs) or Mel frequency cepstral coefficients (MFCCs) are used for developing PRSs. The production of each

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_1

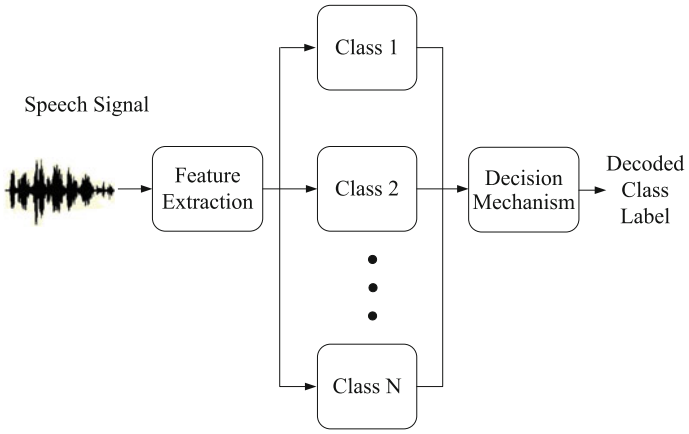

Fig. 1.1 Block diagram of phone recognition system

sound unit is characterized by articulatory and excitation features in addition to vocal tract features. Hence, in this work, the articulatory and excitation source features are explored in addition to spectral features with an intent to improve the performance of PRSs.

The block diagram of PRS is shown in Fig. 1.1. Features are extracted from the input speech signal by the feature extraction block. The extracted features are checked against of N phonetic classes. The decision mechanism block will decode the class label of the input utterance. In this work, HMMs are used for developing PRSs.

1.2 Articulatory Features for Phone Recognition Systems

The articulators such as lips, teeth, tongue, alveolar ridge, hard palate, velum, and glottis are involved in speech production. The articulatory features (AFs) represent the configuration (i.e., positioning) and movement of various articulators during the production of a sound unit. The AFs change from one sound unit to another. AFs can be broadly classified into five groups, namely (i) place, (ii) manner, (iii) roundness, (iv) frontness, and (v) height. The sound units in International Phonetic Alphabet (IPA) chart are arranged based on AFs [9]. The place and manner AF groups capture the characteristics of consonants, while the roundness, frontness, and height AF groups capture the characteristics of vowels. The physical positioning and movements of various articulators can be represented either as continuous values or as discrete values. In this work, the AFs are represented using discrete values. For example, the discrete values for roundness AF group are rounded, unrounded [10]. The significance of having five AF groups to capture various AFs is as follows:

Place of articulation: The air coming out from lungs is obstructed in the vocal tract to produce a sound unit. Different sound units are produced by obstructing airstream in different ways with varying degrees of constriction. *Place of articulation* represents the point of contact between active and passive articulators in the vocal tract, at which obstruction occurs during the production of a consonant. The lower lip and tongue are the typical active articulators, and the remaining articulators represent the passive articulators. For example, the active lower lip comes in contact with passive upper lip to produce a bilabial sound unit. There are eleven different place of articulations.
Manner of articulation: *Manner of articulation* represents the way in which the air escapes from the vocal tract to produce a consonant. For example, the plosive sounds are produced by complete blockage of air followed by a sudden release of air. There are eight different manner of articulations.
Roundedness: *Roundedness* indicates whether the lips are rounded or not, during the production of a vowel.
Frontness: *Frontness* indicates the horizontal position of the tongue relative to the front of the mouth, during the production of a vowel.
Height: *Height* denotes the vertical position of the tongue during the production of a vowel relative to the aperture of the jaw [11].

AFs contain lexical and phonetic information. The speech variability such as coarticulation effect between adjacent sound units is captured by AFs. The AFs act as additional clues, which aid in discriminating between various sound units. AFs provide supplementary information, which can be used along with the spectral features to improve the performance of PRSs.

1.3 Excitation Source Features for Phone Recognition Systems

According the source–filter model of speech production, speech is produced by exciting a linear acoustic filter with an excitation source [12]. The vocal folds form the main source of excitation, and the vocal tract can be viewed as linear acoustic filter. In the speech production system, vocal tract system acts as a time-varying resonator and can be treated as a time-varying filter. The variations in the vocal tract shape can be captured using time-varying filter in the form of resonances and antiresonances of speech spectrum. Just a mere shape of vocal tract without an excitation source would not be sufficient to produce a speech. It is not possible to produce a sound unit without having an appropriate source of excitation. Let us consider bilabial plosive consonants, namely /*p*/ is unvoiced and /*b*/ is voiced. Both /*p*/ and /*b*/ have same place and manner of articulations, but differ only in their excitation type. This means that both /*p*/ and /*b*/ are produced due to same vocal tract shape, but differ in type of excitation. Similarly, there are many other consonants which are produced due to same vocal tract shape but differ in type of

excitation. Similarly, the source of excitation at gross level is same for all vowels, but they differ due to variations in the shape of the vocal tract. Different phones can be distinguished by their unique combination of excitation source and vocal tract shape. The sounds produced due to vibration of the vocal folds are called voiced sounds, and the sounds produced without vibration of the vocal folds are called unvoiced sounds. The periodic opening and closing of the vocal folds result in the harmonic structure in voiced speech signals. The rate of vibration of vocal folds is called the fundamental frequency (F0) of the source of excitation. The periodicity of glottal pulses in the excitation signal can be used for determining the F0. F0 varies from one phone to another phone based on their inherent characteristics as well as the influence of coarticulation characteristics due to the presence of adjacent phones [13]. In [14] and [15], it is reported that the excitation source information derived from linear prediction (LP) residual of the speech signal contains phone-specific characteristics. The characteristics of excitation source features vary from one sound unit to another sound unit. Hence, it is very essential to use the excitation source features in addition to vocal tract features for better discrimination of different classes of phones. Since the excitation source features are robust to the degradations caused by noise [16], excitation source features can be explored for developing robust PRSs.

1.4 Objective and Scope of the Work

Most of the state-of-the-art PRSs are developed using spectral features such as MFCCs or LPCCs. Spectral features mainly represent the gross shape of the vocal tract, but not the information related to the excitation source or the positioning and movements of various articulators. But, the production of each sound unit is characterized by articulatory and excitation features in addition to vocal tract features. A unique combination of articulators in the vocal tract and specific source of excitation results in the production of a particular sound unit. The performance of PRSs can be significantly improved with the use of articulatory and excitation source features along with the spectral features. The primary objective of this book is to improve the performance of PRSs using articulatory and excitation source features in addition to spectral features. The AFs are derived from the spectral features using FFNNs. The evidences obtained from five different AF groups are combined using weighted combination approach to enhance the performance of PRSs. The excitation source features are derived from the LP residual of the speech signal. The excitation source features along with spectral features are used to improve the performance of PRSs. The significance of articulatory and excitation source features is also analyzed for three basic modes of speech, namely read, extempore, and conversation modes.

1.5 Proposed Organization of the Book

- This chapter provides brief introduction about PRS and its applications. Articulatory and excitation source features are described briefly. The objective and scope of the present work is discussed. The chapter-wise organization of the book is provided at the end of this chapter.
- Chapter 2 provides compendious reviews about the use of articulatory and excitation source features to develop speech recognition systems. The literature review for the speech recognition is provided. Various speech features and models used in the context of speech recognition are briefly reviewed in this chapter. Related works using articulatory and excitation source features to improve the performance of speech recognition systems are explained.
- Chapter 3 discusses the proposed approaches to derive and use AFs for phone recognition task. The prediction of AFs for five different AF groups is discussed. The development of tandem and hybrid PRSs using AFs is proposed. The performance of tandem and hybrid PRSs is evaluated, and the results are analyzed.
- Chapter 4 describes the proposed methods to parameterize the excitation source information for phone recognition task. The development of PRSs using combination of spectral and excitation source features is described. The development of robust PRSs using spectral and excitation source features is discussed. The performance of PRSs is evaluated, and the results are compared.
- Chapter 5 explains the development of PRSs using articulatory and excitation source features for read, extempore, and conversation modes of speech. The performance of PRSs across read, extempore, and conversation modes of speech is determined, and the results are analyzed.
- Chapter 6 summarizes the contributions of the present work and provides future directions.

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., S.B. Sunil Kumar, D. Pati, B. Satapathy, K. Sreenivasa Rao, Development of consonant-vowel recognition systems for indian languages: bengali and oriya, *IEEE INDICON* (2013)
4. N.P. Narendra, K. Sreenivasa, Rao, Segment specific concatenation cost for syllable based bengali TTS. Commun. Comput. Inf. Sci. Contemp. Comput. **168**, 371–382 (2011)
5. N.P. Narendra, K. Sreenivasa, Rao, Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis. Appl. Soft Comput. **13**, 773–781 (2013)
6. N.P. Narendra, K. Sreenivasa, Rao, Syllable specific unit selection cost functions for text-to-speech syntheisis. ACM Trans. Speech Language Process. **9** (2012)

7. L. Rabiner, B.-H. Juang, B. Yegnanarayana, *Fundamentals of Speech Recognition* (Prentice-Hall, Upper Saddle River, 2008)
8. P. Richard, Lippmann, Neural network classifiers for speech recognition. The Linc. Lab. J. **1**, 107–124 (1988)
9. The International Phonetic Association, Handbook of the International Phonetic Association, Cambridge University Press. http://www.langsci.ucl.ac.uk/ipa/index.html
10. S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, M. Wester, Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am. **121**, 723–742 (2007)
11. Gerfen, Phonetics Theory. http://www.unc.edu/~gerfen/Ling30Sp2002/phonetics.html
12. G. Fant, Glottal source and excitation analysis. Speech Transm. Lab. Q. Prog. Status Rep. **1**(20), 085–107 (1979)
13. S. Vaseghi, Speech Processing. http://dea.brunel.ac.uk/cmsp/Home_Saeed_Vaseghi/Chapter13-Speech%20Processing.pdf
14. T. G. Csapo, G. Nemeth, A novel codebook-based excitation model for use in speech synthesis, in *IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, (2012) pp. 661–665
15. T. G. Csapo, Increasing the Naturalness of Synthesizes Speech. http://speechlab.tmit.bme.hu/csapo/downloads/Csapo-phonetician2012-paper.pdf
16. B. Yegnanarayana, S.R. Mahadeva, R. Duraiswami, D. Zotkin, Processing of reverberant speech for time-delay estimation. IEEE Trans. Audio Speech Language Process. **13**, 1110–1118 (2005)

Chapter 2
Literature Review

2.1 Introduction

Speech recognition is one of the most active areas of research from last six decades. Many important contributions in speech recognition research are reported in past 25 years. Several researchers have attempted to use articulatory and excitation source features to improve performance of speech recognition systems. There are a very limited number of works using excitation source features for speech recognition, while there are a good number of works exploring the AFs for speech recognition. Few prior works related to the use of articulatory and excitation source features for developing speech recognition systems are briefly discussed in this chapter. The organization of this chapter is as follows : The prior works related to the speech recognition are discussed in Sect. 2.2. Section 2.3 describes the prior works on the development of speech recognition systems using articulatory features (AFs). In Sect. 2.4, the prior works related to the development of speech recognition systems using excitation source features are explained. Section 2.5 summarizes this chapter.

2.2 Prior Works on Speech Recognition

From the existing literature, it is observed that there are lot of works available in the area of speech recognition. This section lists only few important contributions in speech recognition research. In 1959, D.B. Fry [1] presented about the future directions for speech recognition research. He has summarized the working of the human recognition system and described the importance of modeling the speech recognition by machine in a similar way to that of human recognition system. Since the human recognition mechanism depends on both acoustic cues and language characteristics, the language-specific information and spectral features must be combined to improve the performance of speech recognition systems.

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_2

In 1973, Raj Reddy et al. [2] have developed a *HEARSAY* system for voice-chess application. The task of *HEARSAY* system is to recognize a spoken move in a given board position. The model used hypothesis and test paradigm with a set of cooperating independent parallel processes. The information from all the processes is collectively used to recognize the spoken utterance.

In 1976, G.M. White et al. [3] carried out isolated-word recognition using city names and alpha digits. The linear predictive analysis for preprocessing and dynamic programming for classification are used. It is observed that the use of data reduction techniques leads to the reduction in the performance of speech recognition systems.

In 1988, R.P. Lippmann [4] used neural networks for isolated-word recognition. The performance of neural networks is compared with conventional classifiers such as Gaussian and k-nearest neighbor classifiers. The vowel and digit classification experiments are performed. It is observed that neural networks perform better than conventional classifiers for both vowel and digit classification experiments.

In 1989, A. Waibel et al. [5] used time delay neural network (TDNN) for isolated phoneme recognition. The isolated phoneme recognizer was developed using 3 phonemes, namely /b/, /d/, /g/. Three-layered TDNN with error backpropagation is used. The phone recognition accuracy of 98.5% is reported.

In 1989, L.R. Rabiner [6] proposed hidden Markov models (HMMs) for continuous-speech recognition. Three basic problems of HMMs are addressed. Implementation issues related to use HMMs for developing speech recognition systems are explained. The connected-digit and isolated-word recognizers are developed. This is one of the very important contributions to speech recognition research.

In 1989, K.-F. Lee et al. [7] used HMMs for developing a continuous-speech recognizer. TIMIT speech corpus with 39 phones is used. Linear prediction cepstral coefficients (LPCCs) are used as spectral features, and Viterbi decoding was used for decoding the test utterances.

In 1990, F. Fallside et al. [8] have developed continuous-speech recognizer using neural networks. TIMIT corpus with 61 phones is used. The development of phoneme-to-word recognizer is described.

In 1994, H.A. Bourlard et al. [9] proposed hybrid HMM/multilayer perceptron (MLPs) approach for speech recognition. In hybrid HMM/MLP approach, the state emission probabilities of HMMs are estimated using MLPs. Speech recognition systems are developed using HMMs, MLPs, and combination of HMMs/MLPs. TIMIT speech corpus with 61 phones is used. Viterbi decoding is used for decoding test utterances. The performance of hybrid system developed using the combination of HMMs/MLPs is higher compared to other two systems.

In 2000, H. Hermansky et al. [10] proposed the development for tandem speech recognition systems. In tandem speech recognition systems, the output of the first stage is used as feature to develop the second stage. The posterior probabilities obtained from MLPs in the first stage are used as acoustic observations to develop the speech recognition system in the second stage using HMMs. This leads to the combination of discriminative feature processing ability of MLP in the first stage with distribution modeling ability of HMM in the second stage. A reduction of 35%

in the relative error rate compared to conventional Gaussian mixture model (GMM)-HMM-based system is observed.

In 2008, H. Ketabdar et al. [11] proposed a method for more accurate estimation of phone posteriors by the first stage of tandem speech recognition systems. The phone posteriors are better estimated by integrating phonetic and lexical knowledge along with discriminative knowledge. The phonetic and lexical knowledge is captured by using long temporal context. More accurately estimated phone posteriors resulted in the improvement of performance of tandem systems.

Much of the work is not reported in the context of Indian languages. Since the basic units in Indian languages are syllables, the syllable-based speech recognition systems are more appropriate for Indian languages. The syllable is more stable unit than phone as it captures the coarticulation effect well. Few works exploring the syllable-based speech recognition systems for Indian languages are listed below.

In 2004, S.V. Gangashetty et al. [12] have developed syllable-based speech recognition systems for three Indian languages, namely Telugu, Hindi, and Tamil. The syllables are generalized to consonant-vowel (CV) units. The CV units in the continuous speech are spotted using vowel onset points (VOPs) as the anchor points. Support vector machines (SVMs) and autoassociative neural networks (ANNs) are used for developing classification models.

In 2005, S.V. Gangashetty et al. [13] have proposed hybrid HMM/SVM systems by combining the evidences from HMMs and SVMs. The maximum-likelihood estimates of HMMs are combined with the discriminative knowledge captured by SVMs to recognize CV units more accurately. Hybrid HMM/SVM systems have outperformed both HMM-based and SVM-based systems.

In 2012, A.K. Vuppala et al. [14, 15] have proposed two-stage CV recognition system for improving the performance of syllable-based speech recognition system. Two-stage CV recognition system consists of HMMs in the first stage and SVMs in the second stage. HMMs are used for detecting vowel category, while the SVMs are used for detecting consonant category of the CV unit. Telugu broadcast news corpus is used to evaluate the performance of two-stage CV recognition system. It is found that two-stage CV recognition system outperformed the HMM-based and SVM-based single-stage systems. VOP detection methods are discussed in [16, 17]. Syllable-based speech recognition systems are reported in [18].

Few works related to isolated-word recognition systems in the context of Indian languages are listed below. In 2011, K. Kumar et al. [19] have developed isolated-word recognizer for Hindi using HMMs. In 2012, M. Dua et al. [20] have developed an isolated-word recognizer for Punjabi using HMMs.

In recent years, dramatic improvement in the performance of speech recognition systems is achieved by using deep neural networks (DNNs). In 2012, Abdel-rahman Mohamed et al. [21, 22] have used DNNs for speech recognition. DNNs have many layers of hidden units and very large number of parameters. DNNs take coefficients of several frames as input and produce posterior probabilities as output. It is shown that the HMMs with each state modeled using posterior probabilities of DNNs outperform the HMMs with each state modeled using the mixture of Gaussians.

In 2013, A. Graves et al. [23] have explored deep recurrent neural networks for speech recognition. Deep recurrent neural networks involve stacking of multiple recurrent hidden layers on top of each other. The obtained results are comparable with that of DNNs.

In 2013, Tara N. Sainath et al. [24] explored convolutional neural networks (CNNs) for large vocabulary speech recognition (LVCSR). The behavior of features obtained from CNNs is studied for different LVCSR tasks. The behavior of CNNs is compared with DNNs and GMMs. It is found that the CNNs have higher performance compared to DNNs and GMMs. The experiments are conducted using broadcast news corpus and switchboard corpus.

In 2014, Laszlo Toth [25] proposed the use of maxout activation function for CNNs to improve the performance of CNN-based speech recognition systems. It is found that the use of maxout active function resulted in the reduction of phone error rate up to 6%.

In general, speech can be broadly classified into read, extempore, and conversation modes of speech. Read speech involves reading out from the notes such as news reading. Extempore mode of speech is delivered without the aid of notes such as public speaking or delivering a lecture in a class. Conversation mode of speech is an interactive, spontaneous communication between two or more people. More details on read, extempore, and conversation modes of speech are given in Sect. 5.2. All the works described above have used read speech corpus. Few works related to extempore and conversation modes of speech are listed as below.

In 2003, J.L. Gauvain et al. [26] have developed conversational telephone speech recognition system using telephone conversational speech corpus. Speaker normalization and speaker adaptation techniques are employed to improve the performance of conversation speech recognition system.

In 2005, Florian Metze [27] has performed conversational speech recognition. The articulatory features are used to improve the performance of conversational speech recognition systems.

In 2013, Shridhara M V et al. [28] have developed a phone recognition system (PRS) for Kannada language using HMMs. Separate PRSs are developed for read,

Table 2.1 Summary of prior works on speech recognition

• Speech recognition research till 1989, mainly concentrated on the development of isolated-word recognizers
• Development of speech recognition systems using HMMs proposed by Lawrence R. Rabiner in 1989 is one of the major breakthroughs in speech recognition research
• Development of continuous-speech recognition systems started mostly after 1989
• Speech recognition systems are generally developed using HMMs, neural networks, and SVMs
• Tandem and hybrid approaches are most commonly used to improve the performance of speech recognition systems
• State-of-the-art LVCSR systems are developed using CNNs and large amount of training data using DNNs

extempore, and conversation modes of speech, and the results are compared. The phone recognition systems for Bengali and Odia are reported in [29, 30]. A summary of the prior works on speech recognition is provided in Table 2.1.

2.3 Prior Works on Speech Recognition Using Articulatory Features

There are some works exploring the AFs to improve the performance of speech recognition systems. Some of the recent ones are listed as follows: In 2002, Katrin Kirchhoff et al. [31] have used AFs to develop the robust speech recognition systems. The continuous-digit recognition using telephone speech and conversational speech recognition are carried out. It is shown that AF-based systems are capable of achieving superior performance at high noise levels. The combination of acoustic and AFs consistently leads to a significant reduction of word error rate across all acoustic conditions.

In 2005, Florian Metze [27] has used the AFs to improve the performance of conversational speech recognition systems. In 2007, O. Cetin et al. [32] have used AFs to develop the tandem PRSs. The AFs are derived by training MLPs using spectral features. Fisher and switchboard speech corpora are used. The derived AF evidences along with *perceptual linear prediction* features are used to improve the word error rate.

In 2007, Joe Frankel et al. [33] used AFs to develop tandem PRSs. MLP-based AF classifiers are trained using 2000 hours of telephone speech. The recognition accuracies of AF-tandem PRSs are higher than those of phone posterior-based tandem PRSs.

In 2009, Sabato Marco Siniscalchi et al. [34] have used the acoustic-phonetic information to develop speech recognition systems. The acoustic-phonetic information contained the place and manner of articulation. A bank of speech event detectors are used to score place and manner of articulation events using lattice rescoring approach, to derive acoustic-phonetic information. Three tasks, namely continuous-speech recognition, connected-digit recognition, and LVCSR, are carried out. It is found that in all the three cases, systems developed using acoustic-phonetic information have shown higher performance.

In 2013, Vikramjit Mitra et al. [35] have estimated articulatory trajectories from speech signals using neural networks. The articulatory trajectories indicate the place of constriction. The estimated articulatory trajectories are combined with MFCCs to develop LVCSR systems. Results show that the use of articulatory information improves the performance in both clean and noisy environments.

In all of the existing works, the AFs are mostly used as tandem features to improve the recognition accuracy of speech recognition systems. Hence, we have proposed weighted combination approach to combine the evidences derived from five different AF groups. The hybrid PRSs are developed using weighted combination of

Table 2.2 Summary of prior works on speech recognition using articulatory features

• AFs are used for developing robust speech recognition systems
• AFs are mostly used as tandem features to improve the performance of speech recognition systems
• There are no works exploring the AFs to improve the performance of speech recognition systems in the context of Indian languages
• In this book, a weighted combination approach is proposed to combine the evidences derived from different AF groups and AFs are explored in the context of Indian languages using Bengali

various AFs. The systematic analysis of the enhancement of phone-level accuracies contributed by each AF group is carried out. The analysis is carried out by developing separate hybrid PRSs based on the consonant AFs and vowel AFs. From the literature, it is observed that there are no works exploring the AFs to improve the performance of PRSs in the context of Indian languages. Hence, in this book, we have explored AFs in the context of Indian languages using Bengali. Since the AFs provide supplementary information for phone recognition, the combination of spectral and articulatory features may lead to significant improvement in the performance of PRSs. The objective of our study is to use AFs to improve the phone recognition accuracy of PRSs. A summary of the prior works on speech recognition using articulatory features is provided in Table 2.2.

2.4 Prior Works on Speech Recognition Using Excitation Source Features

There are very limited works exploring the excitation source features for speech recognition. Some of the recent works exploring the excitation source features for speech recognition are listed as follows. In 1996, Jialong He et al. [36] have used linear prediction (LP) residual features, containing excitation source information, to improve the performance of isolated-word recognizer. HMM-based speaker-independent isolated-word recognizer is developed using OGI-ISOLET speech corpus. An improvement of 13% was observed in the recognition accuracy. They have concluded that LP residual features contain useful information for speech recognition and act as complementary information to improve the recognition accuracy.

In 1998, Rathinavelu Chengalvarayan [37] has used LP residual features, containing excitation source information, to improve the performance of city name recognizer. A combination of LPCCs and LP residual features leads to the reduction of 8% in the string error rate [37]. In 2008, M. Chetouani et al. [38] claim that LP residual feature contains both linguistic and speaker information.

Table 2.3 Summary of prior works on speech recognition using excitation source features

• Excitation source features are mostly used for improving the performance of isolated-word recognition systems
• There are no works exploring the excitation source features for continuous-speech recognition
• Excitation source features for developing continuous-speech recognition systems are proposed

In 2011, N. Dhananjaya et al. [39] have hypothesized the manner of articulation (MOA) using excitation source information. HMM-based MOA recognizer is developed for five broad MOA categories using TIMIT speech corpus. The acoustic-phonetic information extracted from excitation source features is used to detect and correct the errors at the output of HMM-based MOA recognizer.

In all of the existing works, the excitation source features are mostly used for improving the performance of isolated-word recognition systems. In [36, 37], the excitation source features are used for improving the recognition accuracies of the isolated spoken letter recognizer and the city name recognizer, respectively. There are no works exploring the excitation source features for continuous-speech recognition. Hence, in this book, we have explored excitation source features for developing continuous-speech recognition systems. From the literature, it is observed that there are no works exploring the excitation source features to improve the performance of PRSs in the context of Indian languages. Hence, in this book, we have explored excitation source features in the context of Indian languages using Bengali. The objective of our study is to improve the performance of PRSs using the combination of vocal tract and excitation source features. A summary of the prior works on speech recognition using excitation source features is provided in Table 2.3.

2.5 Summary

In this chapter, overview of prior works on speech recognition and the existing works related to articulatory and excitation source features for developing speech recognition systems are briefly described. There are no works exploring the excitation source features for continuous-speech recognition. Articulatory features are mostly used as tandem features to improve the performance of speech recognition systems. There are no works exploring the articulatory and excitation source features to improve the performance of speech recognition systems in the context of Indian languages. Hence, in this book, articulatory and excitation source features are explored for an Indian language Bengali.

References

1. D.B. Fry, Theoretical aspects of mechanical speech recognition. J. B. Inst. Radio Eng. **19**, 211–218 (1959)
2. D. Raj Reddy, L.D. Erman, R.B. Neely, A model and a system for machine recognition of speech. IEEE Trans. Audio and Electroacoust. **AU-21**, 229–238 (1973)
3. G.M. White, R.B. Neely, Speech Recognition experiments with linear predication, bandpass filtering, and dynamic programming. IEEE Trans. Acoust. Speech Signal Process. **ASSP-24**, 183–188 (1976)
4. R.P. Lippmann, Neural network classifiers for speech recognition. Linc. Lab. J. **1**, 107–124 (1988)
5. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, Phoneme recognition using time-delay neural networks. IEEE Trans. Acoust. Speech Signal Process. **37**, 328–339 (1989)
6. L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989)
7. K.-F. Lee, H.-W. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **37**, 1641–1648 (1989)
8. F. Fallside, H. Lucke, T.P. Marsland, P.J. O Shea, M.S.J. Owen, R.W. Prager, A.J. Robinson, N.H. Russell, Continuous speech recognition for the TIMIT database using neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1990), pp. 445–448
9. H.A. Bourlard, N. Morgan, *Connnectionist Speech Recognition: A Hybrid Approach* (Kluwer Academic Publishers Norwell, USA, 1994)
10. H. Hermansky, D.P.W. Ellis, S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2000), pp. 1635–1638
11. H. Ketabdar, H. Bourlard, Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2008), pp. 4065–4068
12. S.V. Gangashetty, C.C. Sekhar, B. Yegnanarayana, Spotting consonant-vowel units in continuous speech using autoassociative neural networks and support vector machines, in *IEEE Workshop on Machine Learning for Signal Processing* (2004), pp. 401–410
13. S.V. Gangashetty, C.C. Sekhar, B. Yegnanarayana, Combining evidence from multiple classifiers for recognition of consonant-vowel units of speech in multiple languages, in *IEEE International Conference on Intelligent Sensing and Information Processing* (2005), pp. 387–391
14. A.K. Vuppala, K. Sreenivasa Rao, S. Chakrabarti, Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points. Circuits Syst. Signal Process. **31**, 1459–1474 (2012)
15. A.K. Vuppala, K. Sreenivasa Rao, S. Chakrabarti, Improved consonant-vowel recognition for low bit-rate coded speech. Int. J. Adapt. Control Signal Process. **26**, 333–349 (2012)
16. A.K. Vuppala, J. Yadav, K. Sreenivasa Rao, S. Chakrabarti, Vowel onset point detection for low bit rate coded speech. IEEE Trans. Audio Speech Lang. Process. **20**, 1894–1903 (2012)
17. A.K. Vuppala, K. Sreenivasa Rao, S. Chakrabarti, Improved vowel onset point detection using epoch intervals. AEU - Int. J. Electron. Commun. **66**, 697–700 (2012)
18. Manjunath K.E., SBS Kumar, D. Pati, B. Satapathy, K. Sreenivasa Rao, Development of consonant-vowel recognition systems for Indian languages: Bengali and Oriya, in *IEEE INDICON* (2013)
19. K. Kumar, R.K. Aggarwal, Hindi Speech Recognition system using HTK. Int. J. Comput. Bus. Res. **2**, 1–12 (2011)
20. M. Dua, R.K. Aggarwal, V. Kadyan, S. Dua, Punjabi automatic speech recognition using HTK. Int. J. Comput. Sci. Issues **9**, 359–363 (2012)
21. A. Mohamed, G.E. Dahl, G. Hinton, Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**, 14–22 (2012)

22. G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. **29**, 82–97 (2012)
23. A. Graves, A.-R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 6645–6649
24. T.N. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, Deep convolutional neural networks for LVCSR, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 8614–8618
25. L. Toth, Convolutional deep maxout networks for phone recognition, in *International Speech Communication Association (INTERSPEECH)* (2014), pp. 1078–1082
26. J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, F. Lefevre, Conversational telephone speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2003), pp. 212–215
27. F. Metze, Articulatory features for conversational speech recognition. Ph.D. dissertation, Carnegie Mellon University, 2005
28. M.V. Shridhara, B.K Banahatti, L. Narthan, V. Karjigi, R. Kumaraswamy, Development of Kannada speech corpus for prosodically guided phonetic search engine, in *IEEE International Oriental COCOSDA (OCOCOSDA)* (2013), pp. 1–6
29. Manjunath K.E., K. Sreenivasa Rao, D. Pati, Development of phonetic engine for Indian languages: Bengali and Oriya, in *IEEE International Oriental COCOSDA* (2013)
30. 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)
31. K. Kirchhoff, G.A. Fink, G. Sagerer, Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. **37**, 303–319 (2002)
32. O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, K. Livescu, An articulatory feature-based tandem approach and factored observation modeling, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2007), pp. 645–648
33. J. Frankel, M. Magimai-Doss, S. King, K. Livescu, O. Cetin, Articulatory feature classifiers trained on 2000 hours of telephone speech, in *International Speech Communication Association (INTERSPEECH)* (2007), pp. 36–41
34. S.M. Siniscalchi, C.-H. Lee, A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Commun. **51**, 1139–1153 (2009)
35. V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan, M. Liberman, Articulatory trajectories for large-vocabulary speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 7145–7149
36. J. He, L. Liu, G. Palm, On the use of residual cepstrum in speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1996), pp. 5–8
37. R. Chengalvarayan, On the use of normalized LPC error towards better large vocabulary speech recognition systems, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1998), pp. 17–20
38. M. Chetouani, M. Faundez-Zanuy, B. Gas, J.L. Zarader, Investigation on LP-residual representations for speaker identification. Pattern Recognit. **42**, 487–494 (2009)
39. N. Dhananjaya, B. Yegnanarayana, S.V. Gangashetty, Acoustic-phonetic information from excitation source for refining manner hypotheses of a phone recognizer, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2011), pp. 5252–5255

Chapter 3
Articulatory Features for Phone Recognition

3.1 Introduction

In the previous chapter, we have discussed about the existing works related to speech recognition using articulatory and excitation source features. In this chapter, articulatory features (AFs) are explored for improving the performance of the phone recognition systems (PRSs). In this work, AFs are derived from the spectral features using feedforward neural networks (FFNNs) [1]. Mel frequency cepstral coefficients (MFCCs) are used for representing the spectral features. We have considered five AF groups, namely manner, place, roundness, frontness, and height. Five different AF-based tandem PRSs are developed using the combination of MFCCs and AFs derived from FFNNs. Hybrid PRSs are developed by combining the evidences from AF-based tandem PRSs using weighted combination approach. Baseline PRS is developed using hidden Markov models (HMMs) with MFCCs as features. TIMIT and Bengali read speech corpora are considered for developing PRSs. The performance of hybrid PRSs is compared with the baseline PRS and phone posterior (PP)-based tandem PRS. The systematic analysis of phone-level accuracies contributed by each AF group is carried out.

This chapter is organized as follows: Sect. 3.2 describes the speech corpora used in this work. Section 3.3 discusses the different types of feature extraction techniques used in this work. Section 3.4 describes the development of baseline and tandem PRSs. Section 3.5 provides the details of development of hybrid PRSs using weighted combination scheme. Section 3.6 compares the results of the proposed method with that of the existing methods available in the literature. Section 3.7 summarizes the contents of this chapter.

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_3

3.2 Speech Corpora

For developing and analyzing the performance of proposed phone recognition systems, speech corpora of Bengali and English languages are considered. The *Phonetic and Prosodically Rich Transcribed* (PPRT) speech corpus developed at IIT Kharagpur is used for Bengali language, and for English language, well-known TIMIT database is chosen. The details of PPRT and TIMIT speech corpora are discussed in the following subsections.

3.2.1 Bengali Speech Corpus

The *Phonetic and Prosodically Rich Transcribed* Bengali speech corpus developed at IIT Kharagpur is used in this study [2]. The speech corpus contains speech data collected in read, extempore , and conversation modes of speech. The duration of read speech is 1.2 h, while the duration of extempore and conversation speech is 2.5 h each. PPRT speech corpus contains 16 bit precision, 16 kHz speech wave files in three modes of speech. The speech data in all the three modes of speech are transcribed using International Phonetic Alphabet (IPA) chart. IPA provides one symbol for each distinctive sound. IPA contains unique symbols for denoting 59 consonants, 28 vowels, 31 diacritics, and 19 additional signs. The variations in the consonants and vowels are represented using diacritics. The additional signs indicate suprasegmental qualities such as length, tone, stress, and intonation. Although there are about 160 symbols in IPA chart, a particular language can be represented by using very less number of symbols [3]. In our case, we were able to represent speech utterances in Bengali language with 64 IPA symbols plus one *hyphen* used for indicating silence. The speech data is organized in the form of sentences to carry out experiments. The data used for training and testing was from different speakers. For training, around 80% of data was used and remaining 20% of data was used for testing. Table 3.1 shows the number of speakers and the number of sentences used in this study. The details of count of speakers and the count of sentences are shown separately for read, extempore, and conversation modes of speech. First column indicates three modes of speech. Second and third columns show the number of speakers for male and female genders, respectively, while the fourth and fifth columns indicate the count of sentences present in training and testing set, respectively.

3.2.2 TIMIT Speech Corpus

TIMIT speech corpus is a read speech corpus designed for carrying out acoustic-phonetic studies. The corpus was jointly designed by Massachusetts Institute of Technology, SRI International and Texas Instruments. TIMIT is widely used in

Table 3.1 The number of speakers and number of sentences of read, extempore, and conversation modes of Bengali speech corpus

Speech mode	No. of speakers		No. of sentences	
	Male	Female	Training set	Testing set
Read	8	13	687	166
Extempore	7	4	1195	264
Conversation	22	8	1284	310

development and evaluation of automatic speech recognition systems. TIMIT corpus contains 16 bit precision, 16 kHz speech wave files along with time-aligned orthographic, phonetic, and word transcriptions for each utterance. The transcriptions in TIMIT are hand-verified [4]. The training set and core test set, as suggested in TIMIT documentation, are used for training and testing, respectively. The training set contained data from 462 speakers. Each speaker has spoken 10 short sentences of about 3 to 5 s. The complete train set contained 4620 sentences. The core test set involves 24 speakers with 8 sentences from each speaker. Thus, the complete core test set contained 192 sentences.

3.3 Feature Extraction

In this section, feature extraction techniques to derive spectral and articulatory features are discussed. MFCC features are used for representing the spectral features. The AFs are derived from spectral features using FFNNs. The details of extraction of MFCCs and AFs are discussed in the following subsections.

3.3.1 Mel-frequency Cepstral Coefficients

MFCCs capture the gross shape of vocal tract or oral cavity associated with the production of a sound unit. The following procedure is used for extracting MFCCs from the speech signal. The speech signal is divided into frames with a duration of 25 ms [5]. A frame shift of 10 ms is employed for locating the adjacent frames. The blocked frames are Hamming-windowed to reduce the edge effect while taking the discrete Fourier transform (DFT) on the signal. For each frame, cepstral coefficients are computed using Mel-filter bank with 26 Mel filters. The speech is parameterized into 13 MFCCs including 0th cepstral coefficient and their first and second-order derivatives, resulting in a total of 39 components. More details on MFCC features are given in Appendix A.

3.3.2 Extraction of Articulatory Features

In this study, we have considered five AF groups namely place, manner, frontness, roundness, and height. The following subsections describe the details of prediction of AFs using FFNNs.

3.3.2.1 Articulatory Features

The AFs provide crisp representation of each sound unit, in terms of the positioning and movement of various articulators involved in the production of a specific sound unit. AFs vary from one sound unit to another sound unit. Spectral features such as MFCCs capture only the gross shape of the vocal tract, but not the minute variations in the shape of vocal tract. The co-articulation effect between adjacent sound units is captured by AFs. The AFs provide additional clues for discriminating among various sound units. The use of AFs in the development of PRSs can significantly improve the performance of PRSs. In this study, we have considered five AF groups namely place, manner, frontness, roundness, and height. The discrete information about the positioning and movement of articulators with respect to five AF groups is captured. Each AF group along with their possible AF values is shown in Table 3.2. Table 3.2 shows the articulatory feature specification for Bengali and TIMIT datasets. First column indicates the AF group and the cardinality. The cardinality indicates the number of features in an AF group. Second column lists the possible feature values for each AF group. The possible feature values for *manner* AF group are same for both Bengali and TIMIT datasets, while the possible feature values for remaining

Table 3.2 Articulatory feature specification for Bengali and TIMIT datasets

Bengali (Read Speech)	
AF group (Cardinality)	Features
Place (9)	Bilabial, labiodental, alveolar, retroflex, palatal, velar, glottal, vowel, silence
Manner (6)	Plosive, fricative, approximant, nasal, vowel, silence
Roundness (4)	Rounded, unrounded, nil, silence
Frontness (5)	Front, mid, back, nil, silence
Height (6)	High, low, mid-high, mid-low, nil, silence
TIMIT (Read Speech)	
Place (8)	Bilabial, labiodental, alveolar, palatal, velar, glottal, vowel, silence
Manner (6)	Plosive, fricative, approximant, nasal, vowel, silence
Roundness (5)	Rounded, unrounded, diphthong, nil, silence
Frontness (6)	Front, mid, back, diphthong, nil, silence
Height (7)	High, low, mid-high, mid-low, diphthong, nil, silence

AF groups are different for Bengali and TIMIT datasets. Since certain diphthongs in TIMIT dataset are grouped as separate feature, the feature values for $roundness$, $frontness$, and $height$ AF groups are different in Bengali and TIMIT datasets. This is because, deciding the $roundness$, $frontness$, and $height$ feature values for certain diphthongs in TIMIT dataset is ambiguous.

Figures 3.1 and 3.2 show the histogram of occurrences for vowel and consonant AFs, respectively. The $X-axis$ denotes different consonant and vowel AFs, and the $Y-axis$ indicates the number of occurrences for each AF value. In case of vowel AFs, it can be observed that $mid\ value$ of frontness AF group has least number of occurrences, while the $unrounded\ value$ of roundness AF group has highest number of occurrences. In case of consonant AFs, the number of $plosives$ and $alveolars$ is higher, whereas the number of $glottals$ and $labiodentals$ is lower.

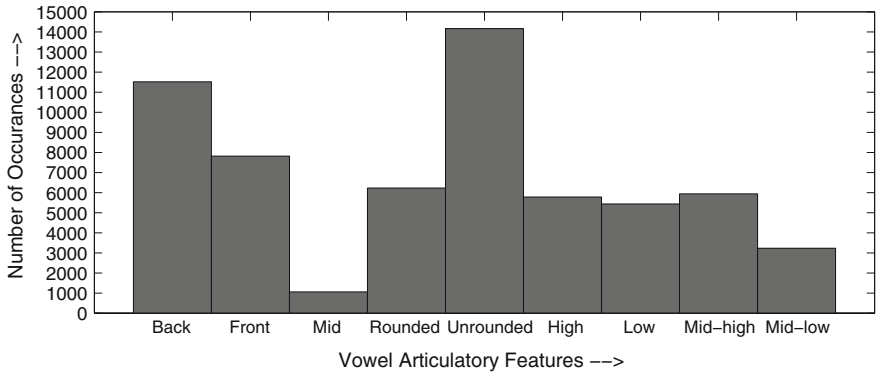

Fig. 3.1 Histogram of occurrences for vowel articulatory features

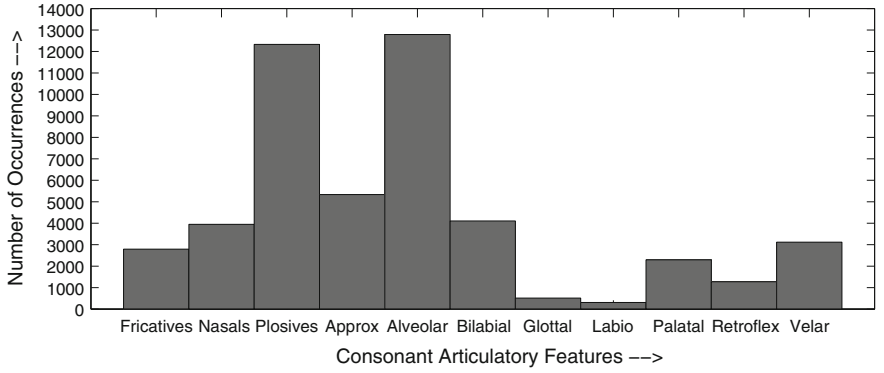

Fig. 3.2 Histogram of occurrences for consonant articulatory features

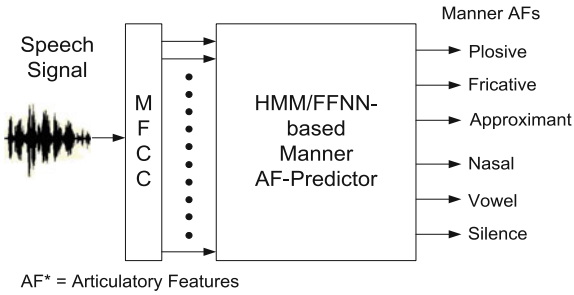

Fig. 3.3 Block diagram of prediction of manner articulatory features

3.3.2.2 Prediction of Articulatory Features

In this work, frame-level AFs for each AF group are predicted from the spectral features using AF-predictors. Separate AF-predictors are developed for each AF group. We have explored both HMMs and FFNNs for developing AF-predictors. Figure 3.3 shows the block diagram of prediction of manner AFs. HMM- and FFNN-based AF-predictors are developed for manner AF group using MFCCs. The predicted feature values represent the manner AFs.

Figure 3.4 illustrates the prediction of manner AFs for ten frames using posteriogram representation. In order to get better visualization of posteriogram distribution across all the feature values, we have plotted the posteriogram using non-consecutive frames. The darker spots in the posteriogram indicate higher posterior probability, while the pale spots indicate lower posterior probability. The labels in the *X-axis* of posteriogram indicate the feature values of manner AF group. MFCCs extracted from each frame are fed to manner AF-predictor to derive the posteriogram distribution for that specific frame. The sum of all the posterior probabilities obtained for a frame will be equal to 1. The posteriogram distribution represents the manner AFs.

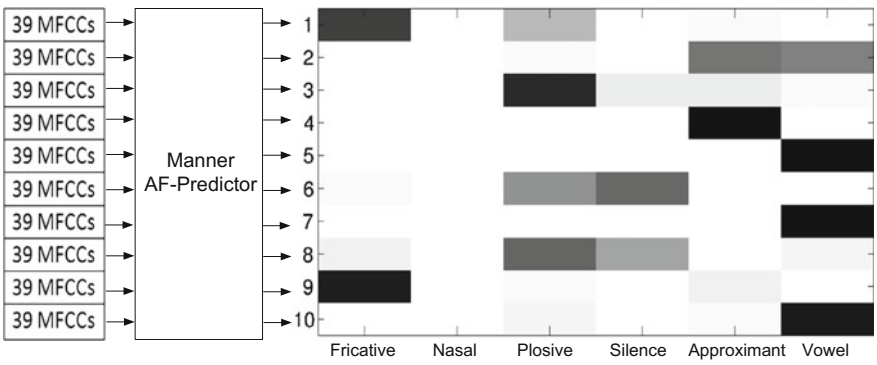

Fig. 3.4 Illustration of prediction of manner articulatory features for ten frames using posteriogram representation

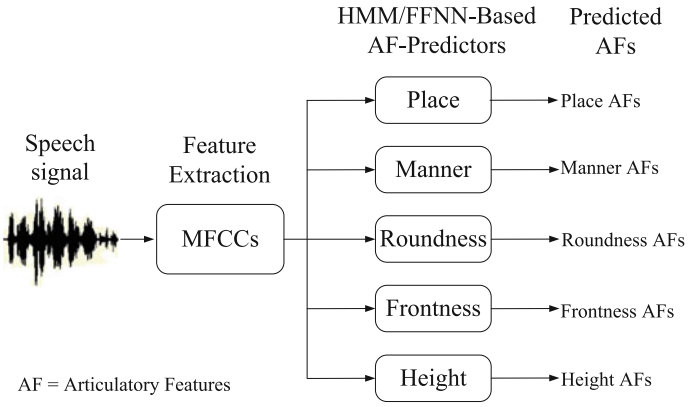

Fig. 3.5 Block diagram of the prediction of articulatory features

Similar kinds of AF-predictors are developed for all five AF groups, as shown in Fig. 3.5. AFs for a particular AF group are predicted using the AF-predictor of that specific group.

Mapping Phone Labels to AF Labels

For training HMMs and FFNNs to develop AF-predictors, we require the speech data which is transcribed at AF level. The AF-level transcription indicates the transcription derived using AF labels. Since the transcription is available at phone level, we derive the AF-level transcription by mapping the phone labels in the phone-level transcription to AF labels. An AF label of an AF group represents a possible AF value for that specific AF group. The possible AF labels for each AF group are shown in Table 3.2. The mapping of each phone label into a set of AF labels of various AF groups for Bengali and TIMIT datasets is shown in Tables 3.3 and 3.4, respectively. First column in Table 3.3 lists unique IPA symbols used in Bengali transcription, while the first column in Table 3.4 lists unique phones used in TIMIT transcription. Second to sixth columns show the corresponding place, manner, roundness, frontness, and height AF values, respectively, for each phone. The mapping for Bengali dataset is derived using IPA chart [3], whereas the mapping for TIMIT dataset is derived with the aid of *TIMIT to IPA mapping* as shown in [6].

Development of AF-Predictors Using HMMs

HMM is a stochastic signal model with finite set of states, and each state is associated with a probability distribution. Transitions among the states are governed by a set of probabilities known as transition probabilities. In a particular state, an outcome or observation can be generated, according to the associated probability distribution.

Table 3.3 Mapping of phone labels to AF groups in Bengali (read speech) dataset

Phones	Articulatory Feature Groups				
	Place	**Manner**	**Roundness**	**Frontness**	**Height**
a	vowel	vowel	unrounded	front	low
o	vowel	vowel	rounded	back	mid-high
ɐ ɜ	vowel	vowel	unrounded	mid	mid-low
i ɪ	vowel	vowel	unrounded	front	high
ɑ	vowel	vowel	unrounded	back	low
ə	vowel	vowel	unrounded	mid	mid-high
ɒ	vowel	vowel	rounded	back	low
u ʊ	vowel	vowel	rounded	back	high
e	vowel	vowel	unrounded	front	mid-high
ɔ	vowel	vowel	rounded	back	mid-low
æ ɛ	vowel	vowel	unrounded	front	mid-low
k k^h g g^h	velar	plosive	nil	nil	nil
ʧ $ʧ^h$ ʤ $ʤ^h$	palatal	plosive	nil	nil	nil
ʈ $ʈ^h$ ɖ $ɖ^h$	retroflex	plosive	nil	nil	nil
t t^h d d^h	alveolar	plosive	nil	nil	nil
p p^h b b^h	bilabial	plosive	nil	nil	nil
m	bilabial	nasal	nil	nil	nil
ɳ	retroflex	nasal	nil	nil	nil
ŋ	velar	nasal	nil	nil	nil
n	alveolar	nasal	nil	nil	nil
s ʃ ʒ	alveolar	fricative	nil	nil	nil
f v	labiodental	fricative	nil	nil	nil
h	glottal	fricative	nil	nil	nil
j	palatal	approximant	nil	nil	nil
ɾ ɹ r l	alveolar	approximant	nil	nil	nil
ɭ	retroflex	approximant	nil	nil	nil
ʋ	labiodental	approximant	nil	nil	nil
sil	silence	silence	silence	silence	silence

In this study, HMM-based systems are developed using a set of context-independent HMMs. A 4-state left-to-right HMM model with a 64 mixture continuous-density diagonal-covariance Gaussian mixture model per state is used to model each sound

Table 3.4 Mapping of phone labels to AF groups in TIMIT dataset

Phones	Articulatory feature groups				
	Place	Manner	Roundness	Frontness	Height
aa	vowel	vowel	unrounded	back	low
ae	vowel	vowel	unrounded	front	low
ah	vowel	vowel	unrounded	back	mid-low
ax ax-h axr	vowel	vowel	unrounded	mid	mid-high
ay	vowel	vowel	unrounded	front	diphthong
eh	vowel	vowel	unrounded	front	mid-low
er	vowel	vowel	unrounded	mid	mid-low
ey	vowel	vowel	unrounded	front	diphthong
ih ix iy	vowel	vowel	unrounded	front	high
uh ux uw	vowel	vowel	rounded	back	high
ow	vowel	vowel	rounded	back	diphthong
ao	vowel	vowel	rounded	back	mid-low
oy aw	vowel	vowel	diphthong	diphthong	diphthong
k kcl g gcl	velar	plosive	nil	nil	nil
t tcl d dcl dx	alveolar	plosive	nil	nil	nil
p pcl b bcl	bilabial	plosive	nil	nil	nil
q	glottal	plosive	nil	nil	nil
th dh s sh	alveolar	fricative	nil	nil	nil
ch jh z zh	palatal	fricative	nil	nil	nil
f v	labiodental	fricative	nil	nil	nil
hh hv	glottal	fricative	nil	nil	nil
l el r	alveolar	approximant	nil	nil	nil
w	labiodental	approximant	nil	nil	nil
y	palatal	approximant	nil	nil	nil
m em	bilabial	nasal	nil	nil	nil
n nx en	alveolar	nasal	nil	nil	nil
ng eng	velar	nasal	nil	nil	nil
epi pau h#	silence	silence	silence	silence	silence

unit. HMMs are trained using maximum likelihood approach. The global *means* and *variances* are computed from the training data to create flat-start HMMs. The embedded re-estimation is carried out on the flat-start HMMs using Baum–Welch algorithm. The number of iterations carried out during re-estimation for Bengali and TIMIT datasets is eight and eleven, respectively. Viterbi decoding is used for finding the hidden sequence of states within a phone, thereby decoding a speech signal into sequence of phones. The open source HTK toolkit is used for building HMM models [7]. More details on HMMs are given in Appendix B.1.

Development of AF-Predictors Using FFNNs

FFNNs are widely explored for developing various speech systems [8–10]. The procedure for developing FFNN-based systems is described in this section. Initially, the frame-level AF labels are assigned for each speech utterance in the training set. For capturing the hidden relations between MFCC features and the AF values of the sound unit, the MFCC feature vectors are given as input and information about AF label is given as output during training of the neural network. The nodes of the network at the input layer have linear functionality, and the nodes at the hidden (second) and output (third) layers have nonlinear functionality. We have experimented with FFNNs of multiple hidden layers, but it was observed that the performance is slightly better using single hidden layer. The lower performance of FFNNs with multiple hidden layers is may be because of insufficient training data. During training, multiple passes are made through the entire set of training data. Each pass is called an epoch. Initially, we start with a learning rate of 0.008. After each epoch, the performance of the FFNNs is measured with a small set of training data, called the cross-validation set, which is held out from main training. The training process will be stopped after the epoch at which the increment in performance improvement is less than 0.5% with cross-validation dataset. The advantage of cross-validation-based adaptive training scheme is that it provides some protection against over-training. The result of training a FFNN is a set of weights. The softmax nonlinearity activation function is used at output layer to constrain posterior probabilities to lie between zero and one and sum to one. The weights associated with the edges between the nodes can then be used as an acoustic model to convert the features of an unseen test utterance into posterior probabilities of each class. The posterior probabilities are used for representing the AFs of a sound unit. The open source quicknet software is used for training FFNNs [11]. Detailed description on FFNNs is given in Appendix B.2.

We have used a memoryless FFNN classifier, which means the outputs depend only on the inputs at that moment. Since the interpretation of the speech sound is highly context-dependent, there is a need to capture the contextual information. The temporal context can be captured by feeding certain frames on either side of the current frame along with the current frame to the input layer. In this study, the temporal context is captured by feeding one frame on either side of the current frame along with the current frame to the input layer. This results in a temporal context of 3 frames with a duration of 45 ms. The number of nodes in input layer (NNIL) is determined using Eq. 3.1.

$$NNIL = No.\ of\ frames\ in\ temporal\ context \times No.\ of\ MFCCs\ per\ frame \tag{3.1}$$

According to Eq. 3.1 the number of nodes in input layer becomes 117, i.e., $3 \times 39 = 117$. The hidden layers with different number of hidden units are tried out. Among all those hidden layers, the hidden layer with 585 hidden units is chosen as a trade-off between computation time required for training FFNNs and performance of the FFNNs. The size of output layer for each AF group is equal to the cardinality of that AF group as shown in Table 3.2. Table 3.5 shows the number of epochs carried out

Table 3.5 Number of Epochs carried out during training of FFNN-based AF-predictors for Bengali and TIMIT datasets

AF Group	Number of Epochs used for training	
	Bengali	TIMIT
Place	10	7
Manner	8	7
Roundness	8	6
Frontness	7	6
Height	9	6

during training the FFNNs for various AF groups of Bengali and TIMIT datasets. First column indicates the AF group. Second and third columns show the number of epochs carried out for Bengali and TIMIT datasets, respectively.

3.3.2.3 Performance Evaluation of AF-Predictors

The accuracy of AF-predictors is determined by comparing the decoded AF labels with the reference transcription of AF labels by performing an optimal string matching using dynamic programming [7]. Once the optimal alignment is found, the number of substitution errors (S), deletion errors (D), and insertion errors (I) is determined. Deletion error indicates that a label is present in the reference transcription but not found in decoded transcription. The substitution error represents that a label in the reference transcription is substituted with some other label in the decoded transcription. The insertion error indicates that a label is present in the is decoded transcription but not found in reference transcription. The recognition accuracy in percentage is calculated using Eq. 3.2.

$$Percentage\ Accuracy\ =\ \frac{N\text{-}D\text{-}S\text{-}I}{N}\ \times\ 100\% \tag{3.2}$$

where N is the total number of labels in the reference transcriptions.

Table 3.6 shows the accuracy of prediction of AFs for different AF groups of Bengali and TIMIT datasets. First column indicates the AF group. Second and third columns show AFs prediction accuracies for Bengali dataset, while the fourth and fifth columns tabulates the AFs prediction accuracies for TIMIT dataset. The results are shown separately for HMM-based and FFNN-based systems. It can be observed that the prediction accuracy of all the AF groups is higher with FFNNs compared to HMMs for Bengali dataset, while the prediction accuracy of most of the AF groups is higher with FFNNs compared to HMMs for TIMIT dataset. Although the prediction accuracies of frontness and height AF groups of TIMIT dataset are higher with HMMs compared to FFNNs, the difference in their prediction accuracies is not significant. Since FFNNs have higher recognition accuracies for all AF groups of Bengali dataset

Table 3.6 Prediction accuracy (%) of AF-Predictors of different AF groups

AF group	Prediction accuracy (%) of AF-Predictors			
	Bengali		TIMIT	
	HMMs	FFNNs	HMMs	FFNNs
Place	55.04	70.35	60.59	67.88
Manner	67.51	74.40	68.47	75.06
Roundness	68.16	78.58	63.13	64.31
Frontness	67.64	74.01	63.00	62.53
Height	62.57	67.75	61.11	60.29

and for majority of AF groups in TIMIT dataset, we have used the FFNNs for predicting the AFs of various AF groups. As FFNNs provide a discriminative way of estimating posterior probabilities [12], it is more advantageous to use FFNNs for developing AF-predictors. The combination of discriminative knowledge captured by AF-predictors and the sequential knowledge captured by HMMs (during the development of PRSs) leads to a kind of hybrid FFNN/HMM system, which has higher potential for improving the recognition accuracies. The following observations are made during the prediction of AFs for different AF groups.

Place: Labiodentals have poor classification accuracy (CA). {labiodental → bilabial, retroflex → alveolar} misclassfications observed. All the groups have significant misclassifications into alveolars. Alveolars and velars have more deletion errors.

Manner: Plosives have very poor CA, which is mainly because of their misclassifications into nasals. Plosives are also misclassified into silence, and this is mostly because of misclassifications of unvoiced plosives such as {*p,t,k*} into silence. Vowels have highest CA.

Roundness: {unrounded → rounded} misclassification is more prominent. Consonants grouped as *nil* are mainly misclassified into vowels and have more deletion errors.

Frontness: *mid* is mainly misclassified to *back* and has got least CA. Consonants grouped as *nil* are mainly misclassified into vowels and have more deletion errors.

Height: {high → mid-high, mid-low → mid-high} misclassifications are prominent. *mid-high* has least CA. Consonants grouped as *nil* are mainly misclassified into *mid-low*.

Tables 3.7, 3.8, 3.9, 3.10, and 3.11 show the confusion matrices for place, manner, roundness, frontness, and height articulatory features (AFs), respectively. The confusion matrices for HMM-based AF-predictors are shown here. Similarly, the confusion matrices could be generated for all the AF-predictors.

Table 3.7 Confusion matrix obtained from HMM-based place AF-predictor (CA = classification accuracy, Labio = labiodental, Retro = retroflex)

	Alveolar	Bilabial	Glottal	Labio	Palatal	Retro	Sil	Vowel	Velar	CA(%)
Alveolar	780	154	52	94	100	219	14	5	52	53.1
Bilabial	11	520	22	50	6	13	3	2	8	81.9
Glottal	2	2	63	3	3	2	0	0	3	80.8
Labio	0	6	2	32	0	1	0	0	0	78.0
Palatal	7	8	8	6	221	13	1	1	1	83.1
Retro	11	12	3	4	1	136	1	0	5	78.6
Sil	0	1	1	0	0	0	384	0	0	99.5
Vowel	11	83	36	140	29	87	21	2449	25	85.0
Velar	12	40	17	34	22	30	7	0	204	55.7

Table 3.8 Confusion matrix obtained from HMM-based manner AF-predictor (CA = classification accuracy)

	Fricative	Nasal	Plosive	Silence	Approximant	Vowel	CA(%)
Fricative	397	22	5	3	5	1	91.7
Nasal	3	548	3	1	26	0	94.3
Plosive	67	231	1373	31	50	1	78.3
Silence	0	0	1	383	1	0	99.5
Approximant	7	115	10	5	471	3	77.1
Vowel	21	104	10	10	161	2600	89.5

Table 3.9 Confusion matrix obtained from HMM-based roundness AF-predictor (CA = classification accuracy)

	Consonant	Rounded	Silence	Unrounded	CA(%)
Consonant	2621	123	85	54	90.9
Rounded	0	931	6	54	93.9
Silence	0	1	389	0	99.7
Unrounded	6	146	7	1826	92.0

Table 3.10 Confusion matrix obtained from HMM-based frontness AF-predictor (CA = classification accuracy)

	Back	Front	Mid	Consonant	Silence	CA(%)
Back	1563	47	185	2	6	86.7
Front	20	988	14	1	3	96.3
Mid	30	2	142	0	0	81.6
Consonant	64	43	148	2662	86	88.6
Silence	0	0	0	0	387	100

Table 3.11 Confusion matrix obtained from HMM-based height AF-predictor (CA = classification accuracy)

	High	Low	Mid-high	Mid-low	Consonant	Silence	CA(%)
High	522	23	60	120	0	6	71.4
Low	5	784	47	49	0	5	88.1
Mid-high	55	70	463	96	1	7	66.9
Mid-low	6	42	25	490	0	3	86.6
Consonant	54	88	60	188	2729	56	86.0
Silence	0	0	1	0	0	386	99.7

3.3.3 Prediction of Phone Posterior Features

PPs are predicted from the spectral features using FFNNs. FFNNs perform the phone classification at frame level. Although HMMs can be used for estimating phone posteriors, FFNNs are employed for this purpose. This is because, FFNNs being discriminative classifiers provide a discriminative way of estimating phone posteriors, while the sequential knowledge capturing ability of HMMs is exploited in later stage of development of PRSs using HMMs. The PPs of phone classes of each frame $p(q_t = i|x_t)$, where q_t is a phone at time t, $i = 1,\ 2 \ldots N$, and x_t is the acoustic feature vector at time t such that

$$\sum_{i=1}^{N} P(i) = 1,$$
$$where\ N = Total\ number\ of\ phone\ classes.$$
$$i\ =\ indicates\ pecific\ phone\ class. \tag{3.3}$$

FFNN is trained, for predicting the PPs, using the procedure mentioned in Sect. 3.3.2.2. The weights associated with the edges between the nodes are used as the acoustic model to convert the features of an unseen test utterance into phone posteriors of each class. Figure 3.6 illustrates the prediction of PPs for ten frames using posteriogram representation. For better visualization of posteriogram distribution across all the phones, posteriogram is plotted using non-consecutive frames. The darker spots in the posteriogram indicate higher posterior probability, while the pale spots indicate lower posterior probability. The labels in the $X\text{-}axis$ of posteriogram indicate the phones used for training the FFNNs. MFCCs extracted from each frame are fed to manner AF-predictor to derive the posteriogram distribution for that specific frame. The sum of all the posterior probabilities obtained for a frame will be equal to 1. The posteriogram distribution represents the PPs. The PPs contain the discriminative knowledge for discriminating between various phonetic units [12].

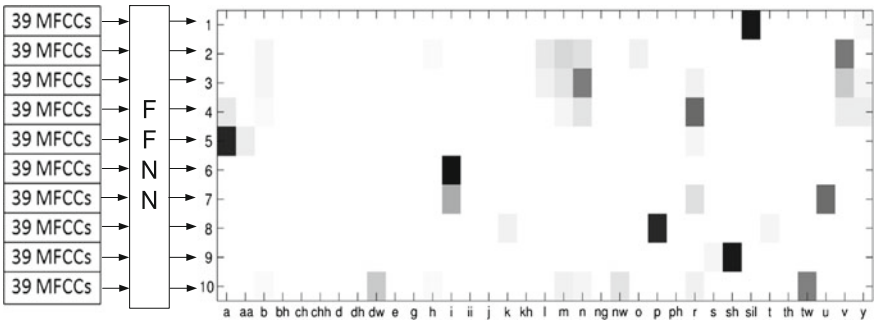

Fig. 3.6 Illustration of prediction of phone posteriors for ten frames using posteriogram representation

The dimension of generated PPs will be equal to the number of phones considered for training FFNNs. We have used a temporal context of 3 frames, which results in a input layer of 117 units. The hidden layer with 585 hidden units is used. The size of output layer is equal to the number of phones considered for training FFNNs [12].

3.4 Development of Baseline and Tandem Phone Recognition Systems

In this study, we have developed Bengali and English PRSs using HMMs. The number of phones considered for developing Bengali and TIMIT PRSs is 35 and 48, respectively. Most frequently occurring phones in the IPA transcription were considered for building Bengali PRS. The 61 phones of TIMIT dataset are downsized to 48 phones by using the approach shown in [13]. HMM-based PRSs are developed using the procedure mention in Sect. 3.3.2.2. The baseline PRSs are developed using MFCCs as features. We have developed AF-based tandem PRSs using combination of MFCCs and the predicted AFs as features. The AFs for each AF group are predicted from the spectral features using the FFNNs, as per the procedure mentioned in Sect. 3.3.2.2. In tandem approach, FFNNs are first trained to perform the classification at frame level, and then, the frame-level posterior probability estimates of the FFNNs are used as the acoustic observations in HMMs. The predicted AFs of a particular AF group are augmented with MFCCs to develop AF-based tandem PRS for that AF group [14]. Separate tandem PRSs are developed using the AFs predicted from each AF group. This leads to the development of five different AF-based tandem PRSs. Figure 3.7 shows the block diagram of manner AF-based tandem PRS. Manner AFs are predicted using manner AF-predictor as shown in Fig. 3.3. The predicted manner AFs are combined with MFCCs to develop HMM-based tandem PRS. Similarly, five different tandem PRSs are developed using the predicted AFs from each AF group.

Figure 3.8 illustrates the manner AF-based tandem PRS for ten frames using posteriogram representation. The MFCCs are augmented with the posteriogram

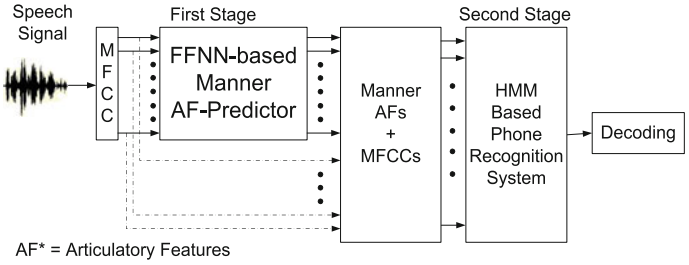

Fig. 3.7 Block diagram of the manner AF-based tandem PRS

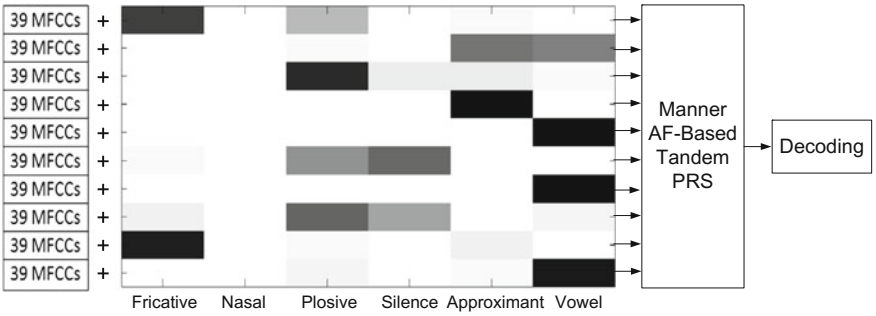

Fig. 3.8 Illustration of manner AF-based tandem PRS for ten frames using posteriogram representation

Table 3.12 Phone recognition accuracy (%) of baseline and AF-based tandem PRSs

Features	Recognition sccuracy (%)	
	Bengali	TIMIT
MFCCs (Baseline)	45.48	58.45
MFCCs + Place AFs	48.89	60.93
MFCCs + Manner AFs	47.74	61.43
MFCCs + Roundness AFs	47.28	60.75
MFCCs + Frontness AFs	46.59	61.11
MFCCs + Height AFs	48.60	61.58

distribution of manner AFs obtained in first stage (shown in Fig. 3.4). The combination of MFCCs and manner AFs is then fed to manner AF-based tandem PRS for decoding the phones in the input speech utterance.

Phone recognition accuracy is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 3.12 shows the phone recognition accuracies of baseline and tandem PRSs for Bengali and TIMIT datasets. First column shows the different types of features used in the development of PRSs. Second and third columns indicate the recognition accuracies of TIMIT and Bengali PRSs, respectively. It can be observed that all tandem PRSs have higher recognition accuracy compared to their respective baseline PRSs. The combination of MFCCs and *place AFs* has shown highest recognition accuracy for Bengali dataset, while the combination of MFCCs and *height AFs* has shown highest recognition accuracy for TIMIT dataset. Among all the vowel AFs, the *height AFs* have shown superior performance for both Bengali and TIMIT datasets.

It is observed that the CA of aspirated plosives is decreased in all the five AF-based tandem PRSs, whereas the CA of most of unaspirated plosives, fricatives, and approximants is increased in all the AF-based tandem PRSs compared to baseline system. The CA of *silence* has improved in all the AF-based tandem PRSs compared to baseline system. The analysis of each AF-based tandem PRS is as follows:

Place AF-based tandem PRS: The CA of nasals and aspirated plosives is decreased, while the CA of all other subgroups is improved. Approximants and nasals have shown the highest and lowest improvements, respectively.

Manner AF-based tandem PRS: The CA of labiodentals is decreased, while the CA of all other subgroups is improved. Vowel and glottal subgroups have shown the highest improvement compared to baseline PRSs.

Roundness AF-based tandem PRS: The CA of both rounded and unrounded vowels is improved. The improvement in the CA of rounded vowels is much higher compared to that of unrounded vowels.

Frontness AF-based tandem PRS: The CA of all the vowels is improved. The back vowels have shown highest improvement in their CAs, while the mid-vowels have shown least improvement in their CAs.

Height AF-based tandem PRS: The CA of all the vowels is improved. The mid-low subgroup has shown least improvement in the CA, while the mid-high subgroup has shown highest improvement in the CA.

3.5 Hybrid Phone Recognition Systems Using Articulatory Features

Hybrid PRSs are developed by combining AF-based tandem PRSs using weighted combination scheme. The performance of Hybrid PRSs is compared with PP-based tandem PRSs. The following subsections describe the details of development and performance evaluation of hybrid PRSs.

3.5.1 Development of Hybrid Phone Recognition Systems Using Articulatory Features

The hybrid PRSs are developed by combining AF-based tandem PRSs using weighted combination approach. In weighted combination scheme, the posterior probabilities from different PRSs are combined at frame level [15]. The combined posterior probability P(j) of each frame with N phone classes, in the test utterance, is given by the Eq. 3.4. The weighting factor w_i varies from 0 to 1 with a step size of 0.1 and sum up to 1 (i.e., $\sum_{i=1}^{k} w_i = 1$).

$$
\begin{aligned}
&For\ each\ frame, \quad P(j) = \sum_{i=1}^{k} w_i * p_i(j), \\
&\quad where, \ \mathbf{j}\ varies\ from\ 1\ to\ N. \\
&\quad N\ = Total\ number\ of\ phone\ classes. \\
&\quad j\ = indicates\ specific\ phone\ class. \\
&\quad k\ = Number\ of\ PRSs\ considered\ for\ combining. \\
&\quad i\ = indicates\ specific\ PRS.
\end{aligned}
\tag{3.4}
$$

Hybrid systems are developed by using the following combinations of AF-based tandem PRSs: (i) place and manner; (ii) roundness, frontness, and height; and (iii) place, manner, roundness, frontness, and height (i.e., all AF-based tandem PRSs). As the place and manner AFs mainly capture the characteristics of consonants, the hybrid PRSs developed using place and manner AF-based tandem PRSs are called consonant-AF-based hybrid PRSs. Since the roundness, frontness, and height AFs mainly capture the characteristics of vowels, the hybrid PRSs developed using roundness, frontness, and height AF-based tandem PRSs are called vowel-AF-based hybrid PRSs. The hybrid PRSs developed using combination of all the five AF-based tandem PRSs are called all-AF-based hybrid PRSs. PP-based tandem PRSs are developed to compare the performance of AF-based hybrid PRSs with PP-based tandem PRSs. The PPs are predicted as per the procedure mentioned in Sect. 3.3.3. The combination of MFCCs and PPs is used for developing PP-based tandem PRSs using HMMs. Figure 3.9 illustrates the PP-based tandem PRS for ten frames using posteriogram representation. The MFCCs are augmented with the posteriogram distribution of PPs obtained in first stage (shown in Fig. 3.6). The combination of MFCCs and PPs is then fed to PP-based tandem PRS for decoding the phones in the input speech utterance.

Figure 3.10 shows the block diagram of development of hybrid PRSs. MFCCs are combined with the predicted AFs of each AF group to develop tandem PRSs for each AF group. The scores from all the five tandem PRSs are combined using weighted

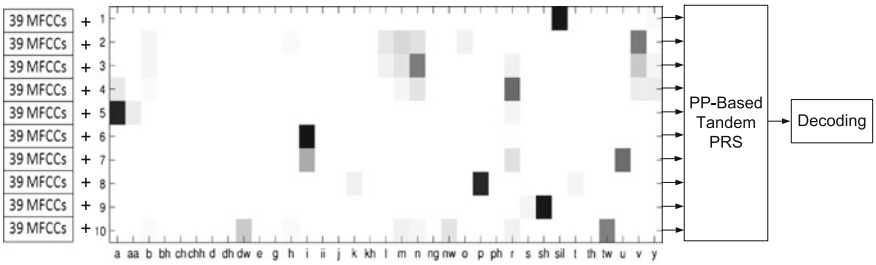

Fig. 3.9 Illustration of PP-based tandem phone recognition system for ten frames using posteriogram representation

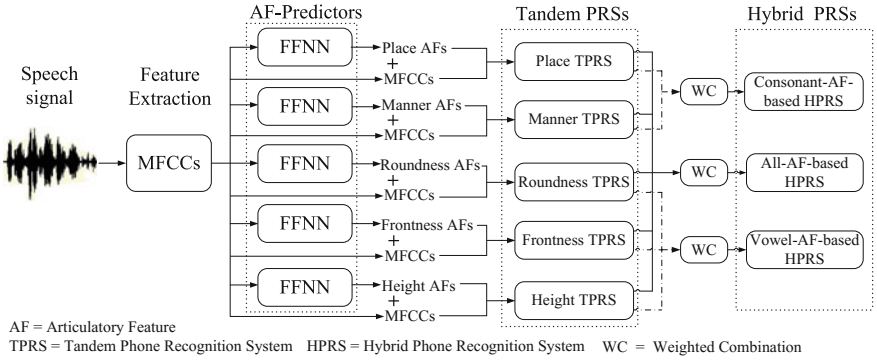

Fig. 3.10 Block diagram of hybrid phone recognition systems

Table 3.13 Weighting factors used for developing Hybrid PRSs using weighted combination approach

Hybrid PRS	Weighting factors									
	Bengali					TIMIT				
	w1	w2	w3	w4	w5	w1	w2	w3	w4	w5
consonant-AF-based	0.5	0.5	–	–	–	0.5	0.5	–	–	–
vowel-AF-based	–	–	0.3	0.3	0.4	–	–	0.3	0.3	0.4
all-AF-based	0.3	0.2	0.2	0.1	0.2	0.3	0.1	0.2	0.1	0.3

combination approach. The scores are combined such that optimal recognition accuracy is achieved.

Table 3.13 shows the optimal weighting factors used for developing hybrid PRSs using weighted combination approach. First column lists the different types of hybrid PRSs. Second to sixth columns indicate the weighting factors for Bengali dataset, while the last five columns indicate the weighting factors for TIMIT dataset. The *hyphen* (-) symbol in Table 3.13 indicates that the particular weighting factor is not applicable for the corresponding hybrid PRS. The weighting factors w1, w2, w3, w4, and w5 correspond to place, manner, roundness, frontness, and height AF-based tandem PRSs, respectively. Among all the combinations of weighting factors considered, the weighting factors listed in Table 3.13 have shown highest recognition accuracies. From Table 3.13, it can be observed that equal weightage is given for both place and manner AF-based tandem PRSs to develop consonant-AF-based hybrid PRSs. In the development of vowel-AF-based hybrid PRSs, a higher weightage is given to the evidence of height AF-based tandem PRSs compared to the evidences from roundness and frontness AF-based tandem PRSs. This is because, the height AF-based tandem PRSs have higher recognition accuracy compared to roundness and frontness AF-based tandem PRSs, as shown in Table 3.12. This is true for both Bengali and TIMIT datasets. The place AF-based tandem PRS, which has

highest recognition accuracy among all the AF-based tandem PRSs, is given highest weightage in the development of Bengali all-AF-based hybrid PRS. Further, we have also combined PP-based tandem PRS and all-AF-based hybrid PRS to develop PP-and-All-AF-based hybrid PRS using the weighting factors 0.3 and 0.7, respectively.

3.5.2 Performance Evaluation of Hybrid Phone Recognition Systems

The phone recognition accuracies of hybrid PRSs is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 3.14 shows the phone recognition accuracies of PP-based and AF-based hybrid PRSs. First column lists the various of PRSs. Second and third columns show the recognition accuracies for Bengali and TIMIT datasets, respectively. It can be observed that the performance of hybrid PRSs is higher than any of the AF-based tandem PRSs. The improvement in the recognition accuracies of hybrid PRSs is consistent; i.e., the recognition accuracy of all-AF-based Hybrid PRSs is higher than both consonant-AF-based and vowel-AF-based hybrid PRSs. Among consonant-AF-based and vowel-AF-based hybrid PRSs, the vowel-AF-based hybrid PRSs have higher recognition accuracies. all-AF-based hybrid PRSs have higher recognition accuracy compared to PP-based tandem PRSs. The PP-and-All-AF-based hybrid PRSs have shown highest recognition accuracy with an improvement of 7.13% and 6.31% for Bengali and TIMIT datasets, respectively, compared to their baseline PRSs.

In all the hybrid PRSs, most of the vowels and unaspirated plosives have shown improvements in their CAs, while most of semivowels, nasals, fricatives, and aspirated plosives have reduction in their CAs. The reduction in the CA of aspirated plosives is mostly because of their misclassification into corresponding unaspirated plosives. The CAs of vowels is more in vowel-AF-based hybrid PRSs compared to that of consonant-AF-based hybrid PRSs, while the CAs of consonants is more in consonant-AF-based hybrid PRSs compared to that of vowel AF-based hybrid PRSs.

Table 3.14 Phone recognition accuracy (%) of PP-based and AF-based hybrid phone recognition systems using Bengali and TIMIT datasets

PRSs using different features	Recognition accuracy (%)	
	Bengali	TIMIT
MFCCs (Baseline)	45.48	58.45
PP-based Tandem PRS	48.97	62.59
consonant-AF-based Hybrid PRS	49.95	61.82
vowel-AF-based Hybrid PRS	51.28	63.04
all-AF-based Hybrid PRS	52.24	63.81
PP-and-All-AF-based Hybrid PRS	52.61	64.76

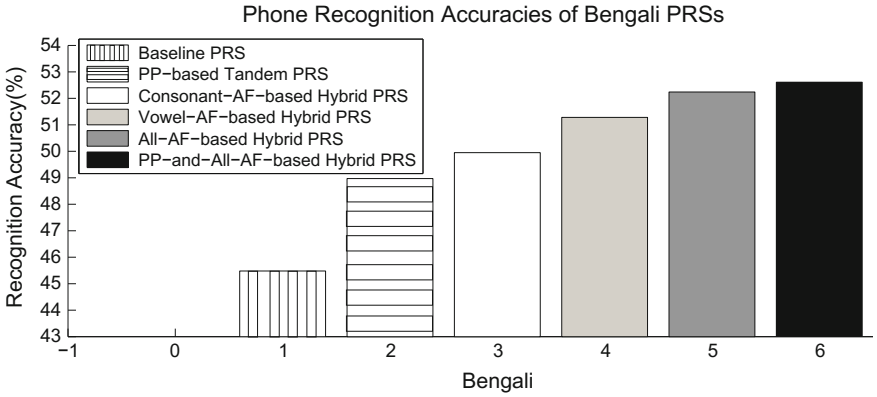

Fig. 3.11 Results of baseline, tandem, and hybrid PRSs plotted using bar graphs for Bengali dataset

Silence has shown improvement in all hybrid PRSs. all-AF-based hybrid PRSs have the CA of vowels, which is in between the CA of vowel-AF-based and consonant-AF-based hybrid PRSs. all-AF-based hybrid PRSs have higher CA of consonants compared to vowel-AF-based and consonant-AF-based hybrid PRSs. This is mainly because of the improvement in the CA of unaspirated plosives. The CA of semivowels is same in both consonant-AF-based and all-AF-based hybrid PRSs. PP-and-All-AF-based hybrid PRSs have highest recognition accuracy in all the subgroups. The improvement in the recognition accuracy of consonants is much higher in PP-and-All-AF-based hybrid PRSs compared to improvements in all other subgroups. The recognition accuracy of semivowels in PP-and-All-AF-based hybrid PRS is almost same as that of baseline PRS.

Figures 3.11 and 3.12 show the results of hybrid PRSs in comparison with the baseline and tandem PRSs plotted using bar graphs for Bengali and TIMIT datasets, respectively. From the Figs. 3.11 and 3.12, it can be observed that the trend in improvement of performance is similar in both Bengali and TIMIT PRSs.

3.6 Discussion of Results

In this work, we have proposed AFs in addition to the well-known spectral features for enhancing the accuracy of the PRSs. From the conducted studies, it is observed that the AFs mainly contribute to resolve the ambiguity in discriminating the phones belonging to certain specific groups. The phones belong to these groups cannot be discriminated by spectral features alone, because all these phones are produced due to similar vocal tract configuration. But, due to the variation in the positioning and movements of various articulators, these phones can be discriminated using appropriate articulatory features.

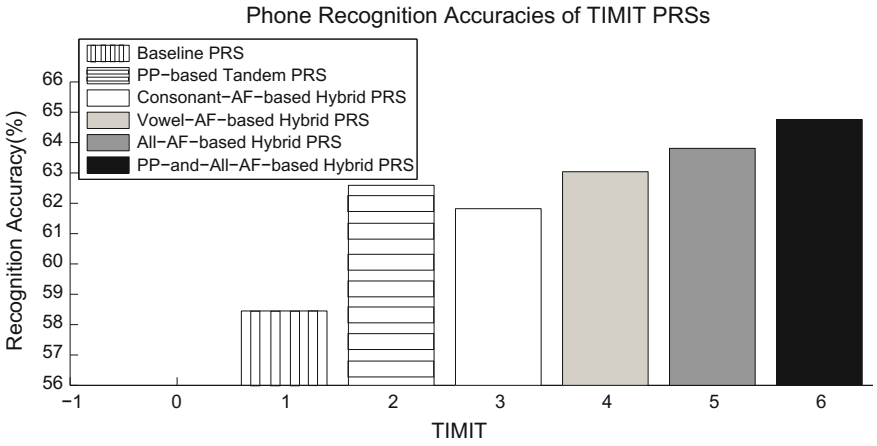

Fig. 3.12 Results of Baseline, tandem, and hybrid PRSs plotted using bar graphs for TIMIT dataset

In this section, we discuss about the performance of the proposed PRSs developed using AFs and compare the results with the state-of-the-art PRSs. In this work, the proposed PRSs are evaluated using Bengali and TIMIT speech databases. The Bengali speech database was developed recently at IIT Kharagpur [2], and hence, we are unable to provide the comparative results of state-of-the-art methods on this database. Even though there exists several works on TIMIT speech database, there are certain difficulties involved in the comparison of the results from different works. Few of these difficulties are listed: (i) The number of phones used for developing and evaluating the phone recognizers is not uniform across the works; (ii) the training and testing sets are not consistent across the works; and (iii) the use of language-related information (i.e., language model) is not consistent across all the works. In the midst of all these difficulties, we have compared the performance of the proposed PRSs with few closely related works and tried to analyze the reasons for either increase or decrease in the recognition accuracy. In order to have consistency in comparison with different works, we have listed all the results in terms of the recognition accuracies. We have expressed all the word error rates and the phone error rates in terms of recognition accuracies.

In 1989, K. Lee et al. have developed context-independent (CI) phone recognizer using HMMs. The training set consists of 2830 sentences from 357 speakers, while the testing set consists of 160 sentences from 20 speakers. It is observed that the phone recognizer developed using 39 phones and without any language model (LM) has a recognition accuracy of 58.77%. In our work, the baseline PRS with 39 phones has shown a recognition accuracy of 63.62%, which is much higher than 58.77% [13]. In 1992, S.J. young has developed a HMM-based phoneme recognizer using the TIMIT dataset. It is found that CI phone recognizer has a recognition accuracy of 52.7% with 39 phones [16]. The training set consists of SI and SX sentences and

testing test contained 160 randomly chosen sentences. Compared to the above work, the performance of our proposed PRS is much better [16, 17].

In 2008, H. Ketabdar et al. have developed a hybrid HMM-/ANN-based phone recognizer using the TIMIT dataset. The standard training and testing sets with 39 phones are used. A long temporal context of 19 frames is used to capture the lexical knowledge. The language-related information is captured using a bigram language model. A recognition accuracy of 71.5% was reported [12]. In this work, we have achieved a recognition accuracy of 71.13%, which is very close to the performance mentioned in the work reported above [12]. The distinction between these two works is due to the following reasons: (i) In [12], the phone recognizer is developed using hybrid HMM/ANN model with a bigram language model (LM), but in our work, phone recognizer is developed using HMMs with no LM; (ii) phone recognition system developed in [12] captures the lexical knowledge using a temporal context of 19 frames, whereas in our work, we have considered the temporal context of 3 frames. In 2009, S. M. Siniscalchi et al. have used the AFs to improve the performance of the HMM-based phone recognizer. A bank of speech event detectors are used to determine the AFs, through a lattice rescoring method. The standard training and testing sets with a set of 45 phones are used. The best obtained result of a CI phone recognizer with no LM has a recognition accuracy of 64.84% [18]. For comparing the results of the above-mentioned system in [18], we have evaluated the proposed PRS with 45 phones, and the recognition accuracy is observed to be 66.78%, which is slightly better compared to the above-mentioned system [18].

In 2011, Dhanajaya et al. have developed HMM-based phone recognizers using the TIMIT dataset. The recognition accuracies obtained using 39 and 48 phones are 61.70% and 56.28%, respectively [19]. In this work, we have achieved recognition accuracies of 63.62% and 58.45% for the baseline PRS using 39 and 48 phones, respectively. By comparing results of [19] with the results obtained in the current work, it can be found that the results obtained in this work are much better compared to that of [19]. In 2011, L. Toth has developed a HMM-based phone recognizer using the TIMIT dataset. The baseline CI phone recognizer with 39 phones has shown a performance of 66.96% [20]. The results obtained from the baseline system of the current study are slightly lower compared to the results of baseline system shown in [20]. The low recognition accuracy by our baseline system may be due to variation in the training dataset used, compared to the PRS developed in [20].

In 2011, R. Rasipuram et al. have used the AFs to improve the performance of the PRSs using TIMIT dataset [21]. The AFs are estimated by training two stages of the multilayer perceptrons (MLPs). First stage takes PLP coefficients as the input and produces AFs as the output. The AFs obtained from the first stage are enhanced by training a second MLP in the second stage. These enhanced AFs along with PPs are used as features to train PRS. The interfeature dependencies between different AF groups are captured using multitask learning (MTL) approach.

Unlike [21], in this work, we have used single-staged MLP to derive AFs. Although the best recognition accuracy reported in [21] is 74.0%, the performance of the PRS developed using the AFs and PPs produced by the first stage is 70.4%. However, the performance of the proposed hybrid PRS is 71.13%, which is better than that of

base-mtl-af+ph PRS in [21]. In [21], AFs are divided into eight AF groups. Naturally, the articulatory posterior probabilities estimated using eight AF groups have more discriminative information compared to the AFs estimated using five AF groups. In [21], 39 phones are used for both training and testing the PRS. But, in our work, we have used 48 phones for training the PRS and 39 phones for testing. The proposed method will produce better or comparable results with that of [21], if we make following changes: (i) using enhanced AFs produced by the second-stage MLP, (ii) use of MTL method to capture the interdependencies between different AF groups, (iii) use of eight AF groups as described in [21], and (iv) use of 39 phones for both training and testing.

In 2013, A. Graves et al. have developed phone recognizer using deep recurrent neural networks (RNNs) and reported a highest recognition accuracy of 82.3% [22]. In 2014, L. Toth has developed a phone recognizer using convolutional deep maxout networks (CDMNs) and obtained a highest recognition accuracy of 83.5% [23]. In 2014, V. Peddinti et al. have developed a phone recognizer using the combination of CNNs and DNNs called CNN/DNN combination networks (CDCNs) [24]. The highest reported recognition accuracy with context-independent HMMs is 81.8%. Although the performance of the proposed method is lesser (i.e., 71.13%) compared to the performance of [22–24], it could be preferred because of the following reasons:

Less Complex: The simplicity of the architecture is measured in terms of number of layers and size of each layer. The architecture of the proposed method is much simpler compared to RNNs, CDMNs, and CDCNs, because of the following reasons:

1. RNNs and CDMNs have 3 and 4 hidden layers, respectively, whereas the proposed approach uses FFNNs with one hidden layer.
2. CDCNs have 2 convolutional layers with 256 hidden units per layer and 4 fully connected layers with 1,024 hidden units per layer. Hence, the architecture of CDCNs is very complex compared to the FFNNs with one hidden layer used in the proposed method.
3. The size of each layer varies between 2714 and 3890 units in CDMNs, which is much higher compared to the 585 units present in the hidden layer of the proposed method.
4. Each input to CDCNs has 2970 parameters. The number of parameters in an hidden layer with 1024 units are 3041280 (around 3 million). The total number of parameters in a network with 2 convolutional and 3 fully connected hidden layers is equal to 10644480, which is very much larger compared to the 68445 parameters used in the proposed hybrid PRS.

Less Training Time: The proposed system requires less training time compared to the RNNs, CDMNs, and CDCNs, because of the following reasons:

Table 3.15 Comparison of the proposed PRSs with the state-of-the-art PRSs on TIMIT database (RA = recognition accuracy, LM = language model, AM = acoustic model, perceptual linear prediction coefficients (PLPCs), * indicates no. of phones)

S.No	Features	Train/Test dataset	Models used	RA (%)	Year	Citation
1	LPCCs	Training: 2830 sentences from 357 speakers, Testing:160 sentences from 20 speakers, No. of phones = 39	AM: HMM, LM: NO	58.77	1989	Lee et al. [13]
2	MFCCs	Training: SI and SX sentences of TIMIT training set, Testing: 160 randomly chosen sentences, No. of phones = 39	AM: HMM LM: NO	52.7	1992	Young [16]
3	PLPCs + Phone posteriors and Temporal context = 19 frames	Training: 3000 utterances from 375 speakers Testing: 1344 utterances from 168 speakers No. of phones = 39	AM: Hybrid HMM/ANN, LM: Bi-gram LM	71.5	2008	Ketabdar et al. [12]
4	MFCCs + AFs	Training: SI and SX sentences of TIMIT training set, Testing: 1344 sentences for testing, No. of phones = 45	AM: HMM, LM: NO	64.84 (45*)	2009	Siniscalchi et al. [18]
5	MFCCs	Training: SA, SI and SX sentences of TIMIT training set, Testing: SA, SI and SX sentences of TIMIT core test set, No. of phones = 39,48	AM: HMM, LM: NO	61.7 (39*) 56.28 (48*)	2011	Dhanajaya et al. [19]
6	MFCCs	Training: SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of TIMIT core test set, No. of phones = 39	AM: HMM, LM: NO	66.96	2011	Toth [20]

(continued)

Table 3.15 (continued)

S.No	Features	Train/Test dataset	Models used	RA (%)	Year	Citation
7	PLPCs + MTL-MLP based AFs produced by first stage + PPs + Temporal context = 17	Training: SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of TIMIT complete test set, No. of phones = 39	AM: KL-HMM, LM: NO	70.40	2011	Rasipuram et al. [21]
8	Log-mel filter bank coefficients, Scatter coefficients and Temporal context of 11 frames	Training: SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of TIMIT core test set, No. of phones = 39	AM: Combination of CNNs and DNNs, LM: NO	81.8	2014	Peddinti et al. [24]
9	Fourier-transform-based Mel-filter bank coefficients	Training: SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of TIMIT core test set, No. of phones = 39	AM: Deep RNNs, LM: NO	82.3	2013	Graves et al. [22]
10	Energy levels of Mel-filter bank channels	Training: SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of TIMIT core test set, No. of phones = 39	AM: hierarchical convolutional maxout networks, LM: Phone bigram LM	83.5	2014	Toth [23]
11	MFCCs	Training: SA, SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of core test set, No. of phones = 39, 45, 48	AM: HMM, LM: NO	63.62 (39*) 66.78 (45*) 58.45 (48*)	2015	Baseline PRS used in our study
12	MFCCs + AFs + Phone posteriors	Training: SA, SI and SX sentences of TIMIT training set, Testing: SI and SX sentences of TIMIT core test set, No. of phones = 39, 48	AM: HMM (Weighted Combination of AF-based Tandem PRSs), LM: NO	71.13 (39*) 64.76 (48*)	2015	Proposed Hybrid PRS

1. Bidirectional RNNs used in [22] process the data in both forward and backward directions. This leads to large number of computations and increases the training time.
2. CDMNs takes 2856 input parameters, which is much higher compared to the 351 input parameters used in the proposed method.
3. The size of the CDCNs input is 2970, which is larger than the input size of 351 used in the proposed system.
4. The computation of scatter features used in [24] requires large number of computations.
5. CDMNs and CDCNs perform operations such as *weight sharing* and *pooling*, which involve significant number of computations, thereby increasing the overall training time.

Less Training Data: The proposed approach works well even with smaller training datasets such as TIMIT, due to smaller size of the input layer. Although CDMNs and CDCNs are trained using TIMIT dataset in [23, 24], respectively, ideally CDMNs and CDCNs should be trained using large training datasets. This is because, CDMNs and CDCNs have large number of input parameters (i.e., large input feature vector), which requires large amount of training data. The use of smaller datasets for training might result in problems such as *curse of dimensionality*.

Table 3.15 summarizes the highlights of the comparative performance of the proposed PRSs with the existing PRSs on the TIMIT dataset.

3.7 Summary

In this chapter, the articulatory features are explored for improving the performance of PRSs. HMM-based Bengali and English PRSs are developed using spectral and articulatory features. The use of articulatory features in addition to spectral features lead to improvement in the performance of PRSs. MFCCs are used as spectral features. AFs are derived from spectral features using FFNNs. Five AF groups, namely (i) place, (ii) manner, (iii) roundness, (iv) frontness, and (v) height, are considered. Five different AF-based tandem PRSs are developed using the AFs predicted from each AF group. Hybrid PRSs are developed by combining the AF-based tandem PRSs using weighted combination approach. all-AF-based hybrid PRSs outperform the conventional PP-based tandem PRSs. all-AF-based hybrid PRSs have higher recognition accuracy compared to consonant-AF-based and vowel-AF-based hybrid PRSs. PP-and-All-AF-based hybrid PRSs developed using combination of all-AF-based hybrid PRSs and PP-based tandem PRSs have shown the highest recognition accuracy. The best obtained results have shown an improvement of 7.13% and 6.31% for Bengali and TIMIT datasets, respectively.

References

1. 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*, pp. 107–111 (2015)
2. S.B. Sunil Kumar, K. Sreenivasa Rao, D. Pati, Phonetic and prosodically rich transcribed speech corpus in indian languages: Bengali and Odia, in *IEEE International Oriental COCOSDA (OCOCOSDA)*, pp. 1–5 (2013)
3. The International Phonetic Association, Handbook of the international phonetic association, Cambridge University Press, http://www.langsci.ucl.ac.uk/ipa/index.html
4. J. Garofolo et al., TIMIT Acoustic-phonetic continuous speech corpus LDC93S1. (Philadelphia: Linguistic Data Consortium, 1993), http://catalog.ldc.upenn.edu/LDC93S1
5. L. Rabiner, B.-H. Juang, B. Yegnanarayana, Fundamentals of Speech Recognition (Pearson Education, 2008)
6. M. Roch, IPA/CMU/TIMIT phone mappings and American English examples, http://roch.sdsu.edu/cs682/IPA-CMU-TIMIT-Phoneset.pdf
7. S. Young et al., The Hidden markov model toolkit and HTK book, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk
8. K. Sreenivasa Rao, Role of Neural network models for developing speech systems, SADHANA, in Academy Proceedings in Engineering Sciences, Indian Academy of Sciences, Vol. 36, Part-5, (Springer, Oct 2011), pp. 783–836
9. A.K. Vuppala, S. Chakrabarti, K. Sreenivasa Rao, Feature mapping using neural network models for coded speech recognition, in *International Conference on Cognitive and Neural systems* (2010)
10. A.K. Vuppala, K. Sreenivasa Rao, Neural network models for speech recognition in mobile environments, in *International Conference on Cognitive and Neural systems* (2009)
11. S. Wegmann et al., QuickNet software and documentation, Speech group at International Computer Science Institute, http://www.icsi.berkeley.edu/icsi/groups/speech
12. H. Ketabdar, H. Bourlard, Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4065–4068 (2008)
13. K.-F. Lee, H.-W. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Sig. Process. **37**, 1641–1648 (1989)
14. H. Hermansky, D.P.W. Ellis, S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1635–1638 (2000)
15. K. Sreenivasa Rao, S.G. Koolagudi, Recognition of emotions from video using acoustic and facial features, in *Signal, Image and Video Processing (SIViP)*, pp. 1–17 (2013)
16. S.J. Young, The general use of tying in phone-based hmm speech recognizers, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I-569–I-572 (1992)
17. C. Lopes, F. Perdigao, Phone recognition on the TIMIT database. Speech Technol. 285–302 (2011)
18. S.M. Siniscalchi, C.-H. Lee, A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Commun. **51**, 1139–1153 (2009)
19. N. Dhananjaya, B. Yegnanarayana, V.G. Suryakanth, Acoustic-phonetic information from excitation source for refining manner hypotheses of a phone recognizer, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5252–5255 (2011)
20. L. Toth, A hierarhical context-dependent neural network architecture for improved phone recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5040–5043 (2011)
21. R. Rasipuram, M. Magimai-Doss, Improving articulatory feature and phoneme recognition using multitask learning. Artif. Neural Netw. Mach. Learn. (ICANN) **6791**, 299–306 (2011)

22. A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013)
23. L. Toth, Convolutional deep maxout networks for phone recognition, in *International Speech Communication Association (INTERSPEECH)*, pp. 1078–1082 (2014)
24. V. Peddinti, T.N. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, V. Goel, Deep scattering spectrum with deep neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 210–214 (2014)

Chapter 4
Excitation Source Features for Phone Recognition

4.1 Introduction

According to the theory of speech production, speech is produced by exciting the vocal tract system with an excitation source [1]. The vocal tract acts as time-varying linear acoustic filter and the vocal folds act as main source of excitation. The state-of-the-art phone recognition systems (PRSs) are mostly developed using vocal tract features, and the excitation source features are not much explored for developing PRSs. Since, the speech is produced by the combination of vocal tract and excitation source characteristics, there is a need for investigating excitation source features in addition to vocal tract system features to improve the performance of PRSs. Hence, in this chapter, we have explored excitation source features to improve the performance of PRSs. The excitation source features have many applications such as language identification [2, 3], emotion recognition [4, 5], speaker verification [6, 7], and speaker recognition [8, 9].

The excitation source information is derived by processing linear prediction (LP) residual of speech signal. The vocal tract information is captured using Mel frequency cepstral coefficients (MFCCs). TIMIT and Bengali read speech corpora are considered for developing PRSs. Further, the excitation source features are explored for developing tandem and robust PRSs.

This chapter is organized as follows: Sect. 4.2 describes the feature extraction techniques used for extracting the excitation source features. In Sect. 4.3, the development of PRSs using vocal tract and excitation source features is discussed. The development of tandem PRSs is explained in Sect. 4.4. Section 4.5 discusses the development of robust PRSs using excitation source features. Section 4.6 summarizes the contents of this chapter.

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_4

4.2 Extraction of Excitation Source Features

In this work, MFCC features are used for capturing vocal tract information, while the residual Mel frequency cepstral coefficients (RMFCCs) and Mel power differences of spectrum in sub-bands (MPDSS) features are considered for capturing excitation source information. MFCC features are extracted as per the procedure mentioned in Sect. 3.3.1. In this section, feature extraction techniques for capturing the excitation source information are discussed. As LP residual mainly contains excitation source information [8, 10–12], in this work, the features derived from LP residual are used to represent excitation source information. Section 4.2.1 describes the procedure for computing LP residual from the speech signal. Sections 4.2.2 and 4.2.3 describe the techniques for parameterizing the excitation source information.

4.2.1 Computation of LP Residual

In LP analysis, the sample $s(n)$ is estimated as a linear weighted sum of the past samples. The predicted sample $\hat{s}(n)$ is given by

$$\hat{s}(n) = -\sum_{k=1}^{P} a_k s(n-k) \tag{4.1}$$

where p is the order of prediction, and $\{a_k\}, \mathrm{k} = 1, 2, \ldots, \mathrm{p}$ is the set of linear prediction coefficients (LPCs). The LPCs are obtained by minimizing the mean-squared error between the predicted sample value and the actual sample value over the analysis frame. The error $e(n)$ between the predicted value $\hat{s}(n)$ and actual value $s(n)$ is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{P} a_k s(n-k) \tag{4.2}$$

This error $e(n)$ is called the LP residual of the speech signal [13]. LP residual signal essentially carries all information that has not been captured by the LP coefficients. LP residual mainly contains excitation source information [8, 10, 11].

4.2.2 Mel Power Differences of Spectrum in Sub-bands

In case of voiced sound units, the rate of vocal folds vibration varies from one sound unit to another. Hence, the periodic information or the harmonic structure of the excitation source also varies from one sound unit to another. The periodicity information in the excitation source can be determined by measuring the difference

between peaks and dips of the LP residual spectrum. The power spectrum $p(k)$ of the LP residual signal $R(k)$ is determined using the relation $p(k) = |R(k)|^2$. The power differences of spectrum in sub-bands (PDSS) features are used for representing the periodicity information present in the excitation source signal [14, 15]. The PDSS is determined using the spectral flatness (SF) measure of the power spectrum in sub-bands. Spectral flatness can be measured by the ratio of geometric mean (GM) to arithmetic mean (AM) of the power spectrum. PDSS of residual sub-band spectra is given by the relation $PDSS = 1 - SF$ and is computed as shown in the Eq. 4.3.

$$V(i) = 1.0 - \frac{\left[\prod_{k=L_i}^{H_i} p(k)\right]^{\frac{1}{N_i}}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} p(k)} \tag{4.3}$$

where $N_i = H_i - L_i + 1$ is the number of frequency points in ith filter. The L_i and H_i are the lower and upper limits of the frequency in ith sub-band, respectively. Since, $0 \leq SF \leq 1$, the values of PDSS also vary from 0 to 1. Higher the periodicity of the LP residual spectrum then the PDSS value is closer to 1.0 and lower the periodicity of the LP residual spectrum then the PDSS value is closer to 0.0. If the spectrum has peaks and dips, i.e., the dynamic range is more, then GM is less than AM and PDSS value is close to one, which implies that the spectrum is more periodic. If the spectrum is nearly flat, i.e., the dynamic range is less, then GM $\simeq$ AM and the PDSS value will be close to zero, which implies that the spectrum is less periodic. So, PDSS measure gives information about the periodicity nature of the spectrum. Sub-band spectra are obtained by multiplying the residual power spectrum with a filter bank. The PDSS values are computed from each sub-band using Eq. 4.3. In this work, the Mel-filter banks are used for computing the PDSS from Mel sub-bands. The motivation for using Mel filters is that the Mel-filter bank is designed based on Mel scale of auditory perception. The Mel-filter bank provides less spectral samples to lower bands and more samples to higher bands (beyond 1 kHz). Since, the Mel filter-bank is used for dividing the power spectrum into sub-bands, the PDSS features, thus obtained are called Mel PDSS (MPDSS) features.

The sequence of operations in deriving MPDSS features is shown in Fig. 4.1. LP residual of the speech signal is obtained by applying inverse filtering on the speech signal as described in Sect. 4.2.1. LP residual signal is windowed into frames with duration of 25 ms with consecutive frame overlap by 10 ms. discrete Fourier transform (DFT) is applied on each frame of LP residual to get LP residual spectrum. The power spectrum is obtained by taking the square of the magnitude of LP residual spectrum. The power spectrum is then passed through a Mel-filter bank with m Mel filters. PDSS for each Mel filter is calculated using Eq. 4.3. The PDSS coefficients from all the Mel filters together represent the MPDSS feature vector. The MPDSS feature vector contains m coefficients.

The distribution of MPDSS features is analyzed by plotting the MPDSS feature distribution of six broad phonetic subgroups. The distribution of MPDSS features of a subgroup is captured by training a Gaussian mixture model (GMM) with 16 mixtures

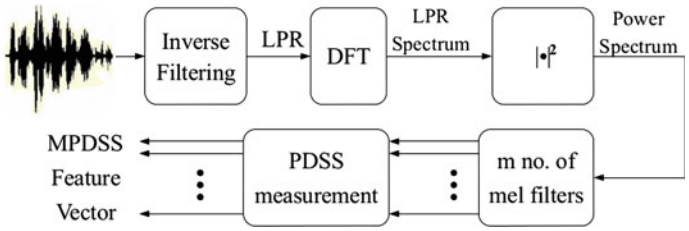

Fig. 4.1 Flow diagram of MPDSS feature extraction

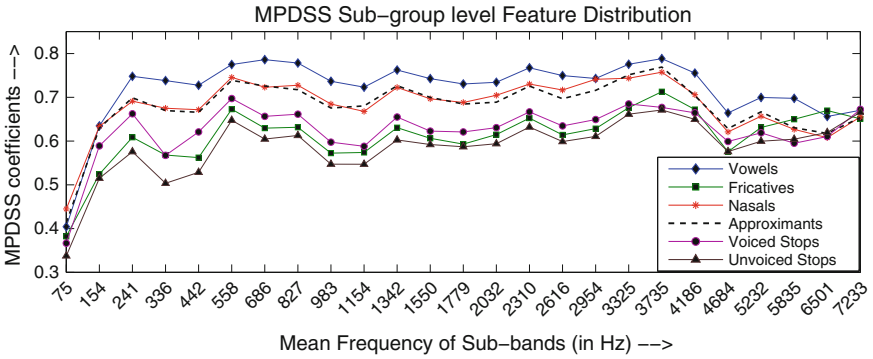

Fig. 4.2 Subgroup-level MPDSS feature distribution

using MPDSS features of that specific subgroup. Separate GMMs are trained using the MPDSS features of each subgroup. The average of the means of 16 mixtures of GMMs is considered for plotting the MPDSS feature distribution for a subgroup. Since, we have considered a Mel-filter bank with 25 Mel filters, each MPDSS feature has 25 coefficients. Figure 4.2 shows the distribution of MPDSS features of different phonetic subgroups. In Fig. 4.2, *X-axis* represents the mean frequency of each sub-band in Hz and the *Y-axis* denotes the MPDSS coefficient value. Mean frequency of a sub-band is obtained by taking the mean of the lowest and highest frequencies of that sub-band. Mean frequency is then rounded off to the nearest integer value.

The six subgroups considered in this work are vowels, nasals, semivowels, voiced stops, fricatives, and unvoiced stops. It can be observed that there is clear separation among the features of all the subgroups. All the voiced groups are at the top, which indicates that their MPDSS value is closer to 1.0 and they have higher periodicity. All the unvoiced groups are at the bottom, which indicates that their MPDSS value is far below 1.0 and they have less periodicity. The feature distributions of nasals and approximants (semivowels) are overlapping. This is because both nasals and approximants are sonorants and both have similar characteristics. The top most plot in Fig. 4.2 corresponds to vowels, which are highly periodic while the bottom most plot corresponds to unvoiced stops, which are aperiodic. The MPDSS features are explored for speaker verification in [6, 7].

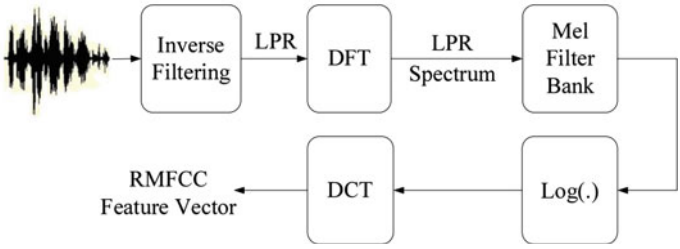

Fig. 4.3 Flow diagram of RMFCC feature extraction

4.2.3 Residual Mel Frequency Cepstral Coefficients

The MFCCs extracted from LP residual signal are called Residual MFCCs (RMFCCs). RMFCCs are used for parameterizing the excitation source information present in the LP residual signal [15, 16]. The sequence of operations in deriving RMFCC features is shown in Fig. 4.3. LP residual of the speech signal is obtained by applying inverse filtering on the speech signal as described in Sect. 4.2.1. LP residual signal is windowed into frames with duration of 25 ms with consecutive frame overlap by 10 ms. DFT is applied on each frame of LP residual to get LP residual spectrum. The LP residual spectrum is then passed through a Mel-filter bank with 26 Mel filters. Discrete cosine transform (DCT) is applied over the Mel-filtered LP residual spectrum to obtain cepstral coefficients. These cepstral coefficients are called RMFCCs, as they are obtained by performing cepstral analysis over LP residual spectrum. In this study, 13 RMFCCs along with their delta and delta–delta coefficients yielding a total of 39 components are considered. The RMFCC features are explored for speaker verification in [6, 7].

4.3 Phone Recognition Systems Using Excitation Source and Vocal Tract System Features

In this study, we have developed Bengali and English PRSs using HMMs. HMM-based PRSs are developed using the procedure mentioned in Sect. 3.3.2.2. The number of phones considered for developing Bengali and TIMIT PRSs are 35 and 48, respectively. Phone recognition accuracy is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 4.1 shows the recognition accuracy of PRSs developed using different types of features for Bengali and TIMIT datasets. First column shows the different types of features used in development of PRSs. Second and third columns indicate the recognition accuracies of TIMIT and Bengali PRSs, respectively.

From Table 4.1, it is observed that the use of excitation source information resulted in improvement of phone recognition accuracy (see 6, 7 and 8 rows). The PRSs developed using excitation source features alone have poor recognition accuracy

Table 4.1 Phone recognition accuracy (%) of PRSs using spectral and excitation source features for Bengali and TIMIT datasets

Features	Recognition accuracy (%)	
	Bengali	TIMIT
MPDSS	11.40	14.93
RMFCC	25.72	35.74
RMFCC + MPDSS	27.30	41.03
MFCCs (Baseline)	45.48	58.45
MFCC + MPDSS	47.29	59.47
MFCC + RMFCC	48.31	60.03
MFCC + RMFCC + MPDSS	48.66	59.53

compared to the PRSs developed using MFCC features (see 2, 3, 4, and 5 rows). This is because MFCCs mainly represent vocal tract information and the vocal tract has major message-bearing articulators, playing a crucial role in production of a sound unit compared to excitation source. The phone recognition accuracy obtained using RMFCC features is higher than the phone recognition accuracy obtained using MPDSS features. This indicates that excitation information is better captured by RMFCC features than MPDSS features. The recognition accuracy obtained using MPDSS alone is least, which indicates that the periodic information captured by MPDSS features alone would not be sufficient to recognize a phone accurately. The combination of RMFCCs and MPDSS has shown higher recognition accuracy than either of RMFCCs or MPDSS features alone. The combination of MFCCs, RMFCCs, and MPDSS has shown highest recognition accuracy for Bengali, whereas the combination of MFCCs and RMFCCs has shown highest recognition accuracy for TIMIT dataset. This shows that the combination of vocal tract and excitation source information helps in better discrimination among different types of phones. The PRSs with highest recognition accuracy have shown an improvement of 3.18% and 1.58% for Bengali and TIMIT datasets, respectively. Though the performance of PRSs developed using excitation source features alone is poor, but it has a good ability to recognize vowels better than other phones [15, 16].

It is observed that, the improvement in the recognition accuracy of combination of MFCCs and excitation source features is mainly because of the improvement in classification accuracies of unaspirated stops. This is because, the excitation source features contain the information for discriminating between aspirated and unaspirated plosive consonants. The strength of excitation is different in aspirated and unaspirated stops. The improvement in classification accuracy of unaspirated plosive consonants is mainly because of the reduction in misclassification of unaspirated plosives to aspirated plosives. There was no improvement in the classification accuracies of nasals and semivowels. The improvement in the classification accuracies of $\{a,e,o\}$ vowels is observed. The classification accuracies of $\{aa,i,u\}$ vowels is reduced with the combination of MFCCs and excitation source features. This is mainly because,

the misclassification exists among the following pairs of vowels: $\{aa \rightarrow a, i \rightarrow e, u \rightarrow o\}$. The use of excitation source features along with MFCCs, resulted in reducing the confusion among the following pairs: $\{k \rightarrow g, g \rightarrow k, j \rightarrow ch, p \rightarrow b, d \rightarrow D, t \rightarrow T\}$. All of the previously mentioned pairs consist of the stops with the same manner and place of articulation, but differing only in their excitation, i.e., voiced or unvoiced phones. Hence, it is clearly observed that the use of excitation source features is responsible for improving the recognition accuracy.

4.4 Tandem Phone Recognition Systems Using Excitation Source and Vocal Tract Features

Most common approach that is used to improve recognition accuracy of PRSs is to develop tandem systems [17]. In this study, we have used the excitation source features to further enhance the performance of tandem PRSs.

4.4.1 Development of Tandem Phone Recognition Systems

The tandem PRSs are composed of two or more stages. In tandem systems, the phone posteriors (PPs) obtained from the first stage will be used as features for developing PRS at the second stage. Generally, the tandem systems are developed using phone posteriors and spectral features such as MFCCs, which mainly represent vocal tract information. Hence, in this study, we have combined phone posteriors with both vocal tract and excitation source information with an intent to improve performance of PRSs. The block diagram of proposed tandem PRS is shown in Fig. 4.4. The PPs are predicted in the first stage using the procedure mentioned in Sect. 3.3.3. The predicted PPs are appended with MFCCs and excitation source features to form the feature vectors for second stage. In the second stage, HMM-based PRSs are developed using various combinations of MFCCs, RMFCCs, MPDSS along with the phone posteriors. The test utterances are decoded using a decoder at the end of second stage.

4.4.2 Performance Evaluation of Tandem Phone Recognition Systems

The recognition accuracy of tandem PRS is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 4.2 shows the recognition accuracy of tandem PRSs developed using Bengali and TIMIT datasets. First column shows the different types of features used in development of tandem PRSs. Second and third columns indicate the recognition accuracies of Bengali and TIMIT tandem PRSs, respectively.

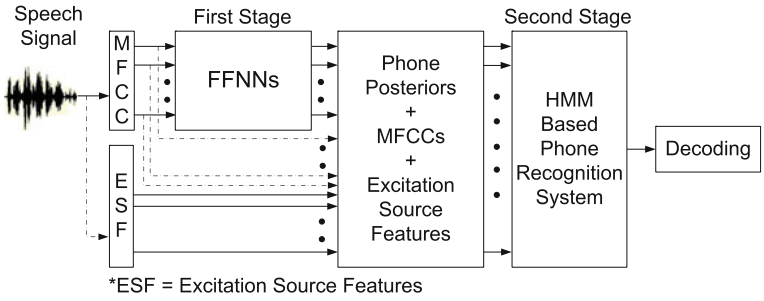

Fig. 4.4 Block diagram of tandem phone recognition system

Table 4.2 Phone recognition accuracy (%) of tandem PRSs (PPs = phone posteriors)

Features	Recognition accuracy (%)	
	Bengali	TIMIT
MFCCs (Baseline)	45.48	58.45
PPs	45.69	59.23
MPDSS + PPs	46.54	60.56
RMFCC + PPs	47.22	61.60
RMFCC + MPDSS + PPs	47.80	62.32
MFCC + PPs	48.97	62.59
MFCC + MPDSS + PPs	49.14	63.04
MFCC + RMFCC + PPs	49.43	63.67
MFCC + RMFCC + MPDSS + PPs	49.57	63.19

From Table 4.2, it can be observed that the phone recognition accuracy of tandem systems is improved by using excitation source features along with MFCCs and PPs. It can be found that the tandem systems developed using PPs alone have higher recognition accuracy than the tandem systems developed using MFCCs alone. The combination of PPs and MFCCs has higher recognition accuracy compared to combination of PPs and excitation source features. This indicates that the vocal tract information contains more phone-specific information. The proposed features, which are the combination of PPs, vocal tract features, and excitation source features, have shown highest performance among all other features. Among RMFCC and MPDSS features, RMFCCs perform better. This is because, in MPDSS only periodicity information is used for discriminating between various phones, whereas the RMFCCs use the spectral information in the LP residual to discriminate among different phones. Spectral information such as weak formant structure present in the LP residual is captured by RMFCCs. The combination of MFCCs, RMFCCs, MPDSS, and PPs has highest recognition accuracy for Bengali dataset, whereas the combination of MFCCs, RMFCCs, and PPs has highest recognition accuracy for

TIMIT dataset. The reduction in the performance of TIMIT dataset with the combination of MFCCs, RMFCCs, MPDSS, and PPs features compared to the combination of MFCCs, RMFCCs, and PPs features might be because of the higher dimensionality of the combination of MFCCs, RMFCCs, MPDSS, and PPs features. Since, the Bengali and TIMIT PRSs are developed using different number of phones, the combination of MFCCs, RMFCCs, MPDSS, and PPs features has lower dimension for Bengali compared to TIMIT dataset, i.e., 138 for Bengali and 151 for TIMIT dataset. The best obtained results have shown an improvement of 4.09% and 5.22% for Bengali and TIMIT datasets, respectively.

The PRSs developed using combination of excitation source features and phone posteriors have higher recognition accuracy compared to PRSs developed using phone posteriors alone. The following observations are made in PRSs developed using combination of excitation source features and phone posteriors in comparison with the PRSs developed using phone posteriors alone. The classification accuracy of fricatives and most of the nasals is reduced. The vowels $\{a,i,u\}$ have shown improvement in their classification accuracies, while the vowels $\{aa,e,o\}$ have shown reduction in their classification accuracies. This is mainly because of the following pairs of misclassifications: $\{aa \rightarrow a, o \rightarrow u, e \rightarrow i\}$. Most of the unaspirated plosives have shown improvement in their classification accuracies, whereas most of the aspirated plosives have shown reduction in their classification accuracies. This is because, the excitation source features contain the information for discriminating between aspirated and unaspirated plosive consonants. The improvement in classification accuracy of unaspirated plosive consonants is mainly because of the reduction in misclassification of unaspirated plosives to aspirated plosives. The PRSs developed using combination of MFCCs and phone posteriors have higher recognition accuracy compared to PRSs developed using the combination of excitation source features and phone posteriors. The overall classification accuracies of vowels, nasals, and fricatives increased in the PRSs using combination of MFCCs and phone posteriors. This is because, the MFCCs have better capability to recognize vowels, nasals, and fricatives compared to excitation source features. The misclassification exists between aspirated and unaspirated plosive consonants. The following observations are made in the PRSs developed using the combination of MFCCs, excitation source features, and phone posteriors in comparison with the PRSs developed using the combination of MFCCs and excitation source features. The classification accuracies of most of the plosives, semivowels, and fricatives are increased, whereas the classification accuracies of most of the nasals are decreased. The vowels $\{a,i,o,u\}$ have shown improvement in their classification accuracies, while the vowels $\{aa,e\}$ have shown reduction in their classification accuracies. This is mainly because of the following pairs of misclassifications: $\{aa \rightarrow a, e \rightarrow i\}$.

4.5 Robust Phone Recognition Systems Using Excitation Source and Vocal Tract Features

In this section, the use of excitation source features to develop robust PRSs is discussed. It is observed that the excitation source features are robust to the degradations caused by noise [18]. Hence, we have attempted to demonstrate the robustness of PRSs using excitation source features [15]. The PRSs are developed using clean speech for TIMIT and Bengali datasets as per the procedure mentioned in Sect. 4.3. Separate PRSs are developed using following features: (i) MFCCs, (ii) MPDSS, (iii) RMFCCs, (iv) combination of MFCCs and MPDSS, and (v) combination of MFCCs and RMFCCs. The recognition accuracy of PRS at a particular signal-to-noise ratio (SNR) is determined by testing the PRS (developed with features from clean speech) using the noisy speech signals of specific SNR. The test utterances are generated by degrading the clean speech signal with additive noises taken from the NOISEX-92 database [19]. The test utterances of various SNRs ranging from 20 to −10 dB are generated.

In this work, white and babble noises are considered to study the robustness of excitation source features. White noise consists of all the frequencies in the audible range 20 Hz to 20 kHz, and it has equal energy at all frequencies and a flat spectrum. The examples of white noise include rain shower, running fan, sound produced by a television or radio when no signal is being received. The addition of white noise affects the speech signal evenly and uniformly at all frequencies regardless of the signal strength distribution in the frequency domain. Babble noise refers to the mixture of speech signals produced by a group of people talking simultaneously. This kind of noise is usually found in large public gatherings. Since the babble noise consists of a continuous low, murmuring sound from multiple speakers, the phone recognition in additive babble noise is challenging. The noises are added to speech signal to generate speech signals with various SNRs. LP residual signal at a particular SNR is obtained by inverse filtering of the noisy speech signal of specified SNR. MFCC features are extracted from the speech signals of various SNRs, while the excitation source features, i.e., MPDSS and RMFCCs, are extracted from the LP residual signals of various SNRs. The features extracted at a particular SNR are used for determining the recognition accuracy at that SNR. The speech signals of different SNRs are generated by using additive white and babble noises.

Figures 4.5 and 4.6 show the subgroup-level MPDSS feature distribution of Bengali language, at different SNRs with additive white and babble noises, respectively. In Figs. 4.5 and 4.6, *X-axis* represent the sub-band number and the *Y-axis* denote the MPDSS coefficient value. The six broad phonetic subgroups considered are: vowels, nasals, semivowels, voiced stops, fricatives, and unvoiced stops. It can be observed that as the noise level increases, the separation between different subgroups decreases. This means that as the noise level increases, the discrimination between phones becomes difficult. The fricatives and voiced stops overlap with each other in almost all the sub-bands, whereas the nasals and approximants overlap with each other only in lower sub-bands. As the noise level increases, the overall value of

MPDSS coefficients decreases, i.e., for 20 dB SNR, the highest MPDSS coefficient is near 0.8 for both noises, whereas for −10 dB SNR the highest MPDSS coefficient of white and babble noises are near 0.6 and 0.65, respectively. This is because, as the noise level increases the harmonic structure in the excitation signal starts degrading and the periodicity information in the speech signal decreases. At −10 dB SNR, the degradation in formant structure is very high in case of white noise, compared to babble noise. Hence, the periodicity information captured by MPDSS is better for the speech signal with additive babble noise than that of additive white noise. This results in more clear separation among the various subgroups of the MPDSS feature distribution of additive babble noise than that of the additive white noise at −10 dB SNR. The subgroups converge toward higher sub-bands in additive white noise, while the subgroups converge toward lower sub-bands in additive babble noise. The separation is more clear in higher sub-bands compared to the lower sub-bands for additive babble noise, whereas it is vice-versa for additive white noise. This is because, the formant structures are clearly visible in the speech signals with additive babble noise even at lower SNRs such as −5 and −10 dB SNRs, whereas the formant structures at lower SNRs are not visible in the speech signals with additive white noise. Hence, the harmonic structure captured by MPDSS features is better in case of additive babble noise than additive white noise. The babble noise has more energy in lower frequencies compared to the energy in higher frequencies. Hence, the addition of babble noise results in the speech signals with more degradations in their lower frequencies compared to their higher frequencies. This leads to higher MPDSS coefficient values at higher sub-bands compared to lower sub-bands, as depicted in Fig. 4.6.

Table 4.3 shows the phone recognition accuracy of PRSs using additive white and babble noises for Bengali and TIMIT datasets. First column shows the SNRs. The second to sixth columns indicate recognition accuracies for Bengali dataset, whereas the last five columns show the recognition accuracies for TIMIT dataset. Second column indicates the recognition accuracy obtained using MFCCs features. The third and fourth columns show the recognition accuracy obtained using MPDSS and RMFCCs, respectively, while the fifth and sixth columns indicate the recognition accuracy obtained using the combination of MFCCs and excitation source features. Similarly, last five columns give the phone recognition accuracies of TIMIT PRSs. It is observed that the combination of MFCCs and excitation source features has superior performance in almost all the cases. The recognition accuracy obtained using excitation source features alone is less than that of MFCCs alone. Although the performance of MPDSS features is lowest in almost all the cases, but the degradation in the recognition accuracy of MPDSS features is not very drastic as compared to all other features. The degradation in the performance of RMFCC features is drastic at higher SNRs, but as the noise level increases, the degradation becomes slow. This indicates that the RMFCC features are more robust to higher noise levels.

Figure 4.7 shows the plot of variation of phone recognition accuracies with respect to different SNRs using additive white noise. In Fig. 4.7, *X-axis* refers to the SNRs in dB and *Y-axis* denotes the phone recognition accuracies in percentage. The following observations are made on the PRSs with additive white noise. In majority of the

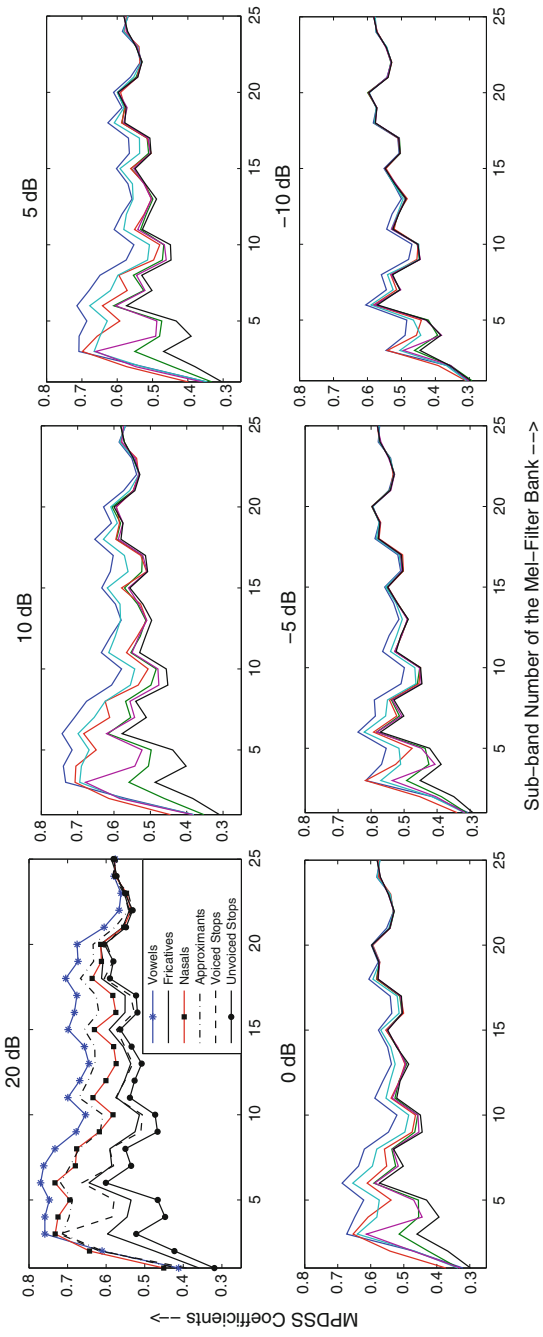

Fig. 4.5 Distribution of MPDSS features for different subgroups at various SNRs with additive white noise

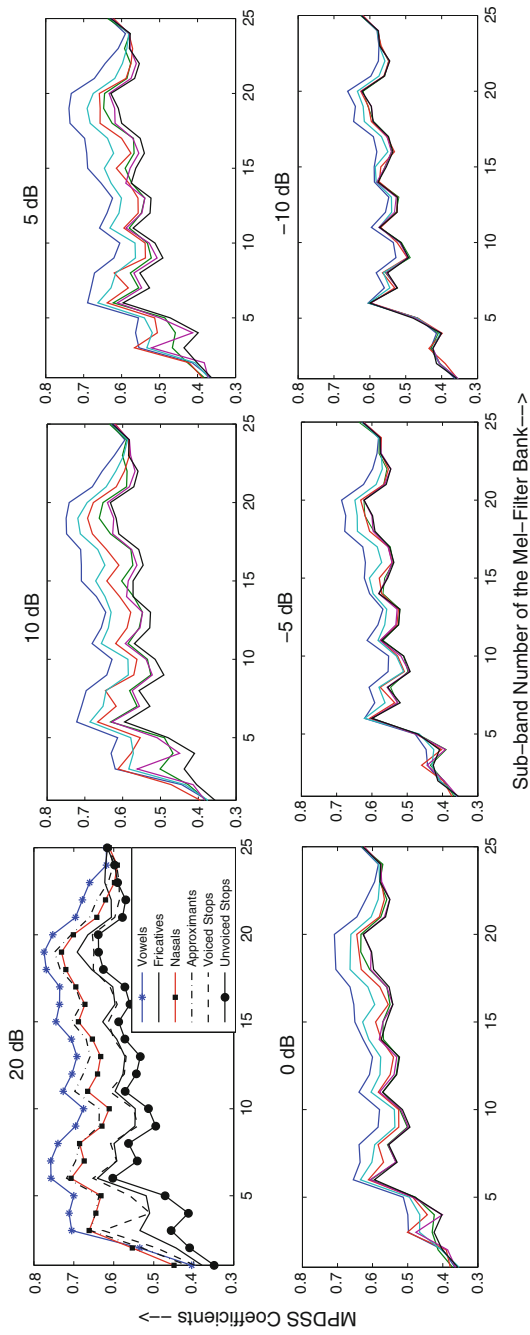

Fig. 4.6 Distribution of MPDSS features for different subgroups at various SNRs with additive babble noise

Table 4.3 Phone recognition accuracy (%) of robust PRSs using additive white and babble noises (MF = MFCC, MP = MPDSS, RM = RMFCC, SNR in dB)

White noise										
SNR	Bengali					TIMIT				
	MF	MP	RM	MF + MP	MF + RM	MF	MP	RM	MF + MP	MF + RM
Clean	45.48	11.4	25.72	47.29	48.31	58.45	14.93	35.74	59.47	60.03
20	33.00	9.37	22.34	37.09	38.81	36.62	9.57	9.66	37.79	41.28
10	23.32	7.59	16.63	26.45	26.39	22.92	6.59	8.37	23.40	24.85
5	19.00	6.35	11.43	22.03	21.57	17.76	5.09	6.65	17.51	18.93
0	16.39	4.92	5.64	18.57	18.76	13.43	3.85	6.05	11.24	13.06
−5	13.28	3.13	2.72	11.94	14.23	9.18	2.55	4.26	4.34	7.31
−10	8.66	2.00	2.02	3.97	6.18	4.09	1.60	3.31	2.69	3.15
Babble noise										
Clean	45.48	11.4	25.72	47.29	48.31	58.45	14.93	35.74	59.47	60.03
20	40.66	10.88	23.65	43.83	43.63	47.70	13.13	10.07	50.05	49.71
10	30.65	10.26	13.97	34.63	34.01	33.64	9.85	8.92	36.67	36.81
5	23.90	9.34	8.56	27.89	26.28	25.09	7.75	8.35	27.72	27.56
0	17.91	6.86	4.97	21.87	21.99	18.34	6.31	7.27	18.61	20.20
−5	13.75	5.14	3.16	16.54	18.32	11.78	4.84	6.23	11.09	15.19
−10	10.60	4.13	2.37	12.89	14.35	8.63	4.15	5.42	5.89	11.86

cases, the combination of MFCCs and excitation source features perform better than the MFCCs alone. In the lower SNRs, i.e., −5 and −10 dB SNRs, the performance of all the three features is almost same. At −10 dB SNR, the MFCCs perform better than the combination of MFCC and excitation source features. This is because, at −10 dB SNR, there would not be much periodicity information present in the excitation source features. In Fig. 4.5, MPDSS feature distribution of various subgroups at −10 dB SNR overlaps with each other, almost leading to a single line. Hence, the discrimination among the various subgroups is very less at −10 dB SNR. Among the two excitation source features, RMFCC features perform better than MPDSS features. The combination of MFCCs and MPDSS features has least phone recognition accuracies at −5 and −10 dB SNRs for Bengali and at 0, −5, and −10 dB SNRs for TIMIT. This is because, at 0, −5, and −10 dB SNRs, the level of white noise is higher than or equal to the level of signal and the spectral flatness of the white noise is close to 1.0. From Eq. 4.3, it can be computed that when the SF is close to 1.0 then MPDSS becomes close to 0.0, indicating a lower periodicity. Hence, the performance of PRSs below 0 dB SNR is less.

Figure 4.8 shows the plot of variation of phone recognition accuracies with respect to different SNRs using additive babble noise. In Fig. 4.8, *X-axis* refers to the SNRs in dB and *Y-axis* denotes the phone recognition accuracies in percentage. The following observations are made on the PRSs with additive babble noise. In case of TIMIT, the

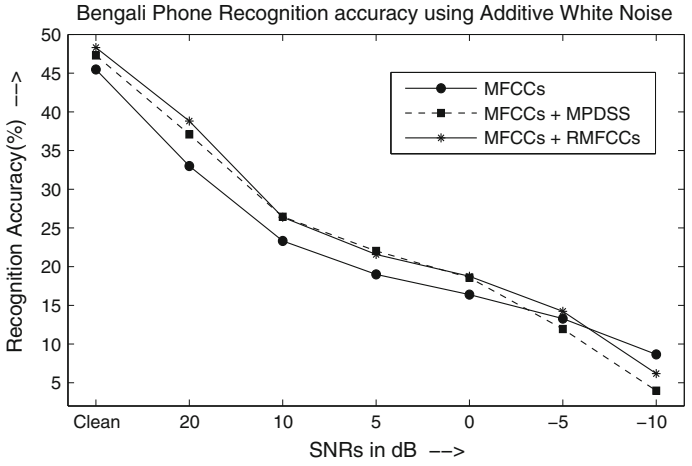

Fig. 4.7 Recognition accuracies of Bengali PRSs at different SNRs with additive white noise (*X-axis* = SNRs in dB and *Y-axis* = phone recognition accuracies (%))

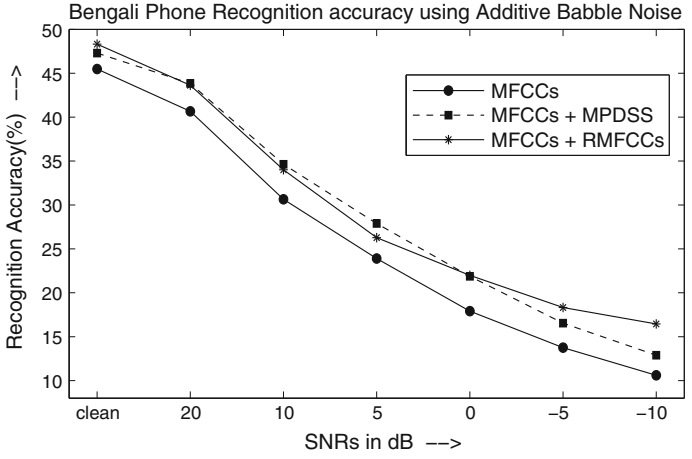

Fig. 4.8 Recognition accuracies of Bengali PRSs at different SNRs with additive babble noise (*X-axis* = SNRs in dB and *Y-axis* = phone recognition accuracies (%))

performance of combination of MFCCs and MPDSS features degrades at −5 and −10 dB SNRs. In Bengali PRSs, the combination of MFCCs and excitation source features has shown highest performance in all SNRs for additive babble noise, while the additive white noise has shown highest performance only in higher SNRs. This is because, the separation among various subgroups at lower SNRs is much better in case additive babble noise compared to additive white noise. Among the two excitation source features, the RMFCC features perform better at lower SNRs, while the MPDSS features perform better at higher SNRs.

4.6 Summary

In this chapter, the excitation source features are explored for improving performance of PRSs. HMM-based PRSs are developed using vocal tract and excitation source features using Bengali and TIMIT datasets. MFCCs are used as vocal tract features, while the RMFCCs and MPDSS are used as excitation source features. The use of excitation source information in addition to vocal tract information has improved the performance of PRSs. As the vocal tract has major role in speech production, the PRSs developed using only vocal tract information have higher recognition accuracy compared to the PRSs developed using excitation source information alone. The use of excitation source features has lead to further improvement in the performance of tandem PRSs. The tandem PRSs developed using the combination of phone posteriors, excitation source information, and vocal tract information have shown highest performance. We have also explored excitation source features to develop robust PRSs. The robustness of PRSs is improved using excitation source information in addition to the vocal tract information. The performance of PRSs is higher in case of additive babble noise compared to additive white noise.

References

1. G. Fant, Glottal source and excitation analysis. Speech Transm. Lab. Q. Prog. Status, Rep. **1**(20), 085–107 (1979)
2. D. Nandi, D. Pati, K.S. Rao, Parametric representation of excitation source information for language identification. Comput. Speech Lang. **41**, 88–115 (2017)
3. D. Nandi, D. Pati, K.S. Rao, Implicit excitation source features for robust language identification. Int. J. Speech Technol. **18** 459–477 (2015)
4. S.G. Koolagudi, S. Devliyal, B. Chavla, A. Barthwal, K.S. Rao, Recognition of emotions from speech using excitation source features, in *International Conference on Modeling Optimization and Computing* (2012)
5. K.S. Rao, S.G. Koolagudi, Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol. **16**, 181–201 (2013)
6. D. Pati, S.R.M. Prasanna, A comparative study of explicit and implicit modelling of subsegmental speaker-specific excitation source information. Sadhana **38**, 591–620 (2013)
7. D. Pati, S.R.M. Prasanna, Speaker verification using excitation source information. Int. J. Speech Technol. **15**, 241–257 (2012)
8. K.S.R. Murty, B. Yegnanarayana, Combining evidence From residual phase and MFCC features for speaker recognition. IEEE Signal Process. Lett. **13**, 52–55 (2006)
9. D. Nandi, D. Pati, K.S. Rao, Robustness of excitation source information for language independent speaker recognition, in *International Oriental COCOSDA Conference* (2013)
10. S.R.M. Prasanna, C.S. Gupta, B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. **48**, 1243–1261 (2006)
11. D. Pati, S.R.M. Prasanna, Non-parametric vector quantization of excitation source information for speaker recognition, in *IEEE Region 10 Conference TENCON* (2008)
12. D. Nandi, D. Pati, K.S. Rao, Sub-segmental, segmental and supra-segmental analysis of linear prediction residual signal for language identification, *IEEE International Conference on Signal Processing and Communications* (2014)
13. J. Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**, 561–580 (1975)

14. S. Hayakawa, K. Takeda, F. Itakura, Speaker identification using harmonic structure of LP-residual spectrum. Biometric Pers. Aunthentification Lect. notes **1206**, 253–260 (1997)
15. Manjunath K.E., K.S. Rao, Articulatory and excitation source features for speech recognition in read, extempore and conversation modes. Int. J. Speech Technol. 1–14 (2015)
16. Manjunath K.E., K.S. Rao, M.G. 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
17. H. Hermansky, D.P.W. Ellis, S. Sharma, Tandem connectionist feature extraction for conventional HMM systems, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2000), pp. 1635–1638
18. B. Yegnanarayana, S.R.M. Prasanna, R. Duraiswami, D. Zotkin, Processing of reverberant speech for time-delay estimation. IEEE Trans. Audio Speech Lang. Process. **13**, 1110–1118 (2005)
19. A. Varga, H.J. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**, 247–251 (1993)

Chapter 5
Articulatory and Excitation Source Features for Phone Recognition in Read, Extempore and Conversation Modes of Speech

5.1 Introduction

In previous two chapters, the articulatory and excitation source features are explored to improve the performance of phone recognition systems (PRSs) developed using read speech corpora. In this chapter, the articulatory and excitation source features are explored using extempore and conversation modes of speech. The performance of PRSs across read, extempore, and conversation modes of speech is compared, and the results are analyzed. This chapter is organized as follows: Sect. 5.2 briefly describes the different modes of speech. Section 5.3 discusses the feature extraction techniques used for extracting the articulatory and excitation source features across three modes of speech. In Sect. 5.4, the development of tandem PRSs using articulatory features is described. Section 5.5 discusses the development of hybrid PRSs for read, extempore, and conversation modes of speech. The development of PRSs for three modes of speech using excitation source features is explained in Sect. 5.6. The analysis of results across read, extempore, and conversation modes of speech is carried out in Sect. 5.7. Section 5.8 summarizes the contents of this chapter.

5.2 Different Modes of Speech

In general, speech can be broadly classified into read, extempore, and conversation modes of speech. The significance of classification of speech into three modes of speech is as follows:

- **Read speech**: Read speech involves reading out from the notes such as television news reading. It is a highly constrained mode of speech, where the message content is made available to the speaker prior. It is more structured, planned, and prepared well in advance. Read speech is delivered using more formal language, and it is one-sided. The speaker prosody variations are minimal in read speech.

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_5

- **Extempore speech**: Extempore speech is delivered without the aid of notes. The subject speaks with confidence and in a bold fashion. The speaker attempts to create an atmosphere to capture the attention of listeners. Delivering a lecture to students in a class is an example of extempore speech. It is more vigorous, flexible, and spontaneous. The extempore mode of speech is also called lecture mode of speech. The prosody usually varies within a limited set of constraints.
- **Conversation speech**: The conversation mode of speech is a form of interactive, spontaneous communication between two or more people, who are following the rules of etiquette. Conversation speech is spontaneous because the conversation proceeds unpredictably. It is informal, unstructured, and unorganized. Conversation speech involves free speaking style with no constraints. In conversation mode of speech, both message and prosody are free from constraints.

It is essential to build separate PRSs for each mode of speech to perform the phone recognition across three modes of speech in an efficient way. Hence, in this chapter, we have analyzed the articulatory and excitation source features in three modes of speech. The articulatory and excitation source features are used for improving the performance of the PRSs in three modes of speech. The performance of PRSs across read, extempore, and conversation modes of speech is evaluated, and the results are analyzed. Bengali speech corpus with speech data in read, extempore, and conversation modes of speech is considered. The details of the speech corpus are given in Sect. 3.2.1.

5.3 Feature Extraction

In this section, the feature extraction techniques to derive the articulatory and excitation source features are discussed. Mel frequency cepstral coefficients (MFCCs) are used for representing the spectral features. Residual Mel frequency cepstral coefficients (RMFCCs) and Mel power differences of spectrum in sub-bands (MPDSS) features are used for capturing the excitation source information. MFCC features are extracted as per the procedure mentioned in Sect. 3.3.1 [1, 2]. The excitation source features are extracted using the procedure mentioned in Sect. 4.2. The AFs are extracted using the procedure given in Sect. 3.3.2 [3, 4]. The specific details of extracting the articulatory features (AFs) from extempore and conversation modes of speech are described in the following subsections.

5.3.1 Articulatory Features for Extempore and Conversation Modes of Speech

Table 5.1 shows the articulatory feature specification for extempore and conversation modes of speech. First column indicates the AF group and the cardinality. The

Table 5.1 Articulatory feature specification for extempore and conversation modes of speech

Extempore and conversation modes of speech (Bengali)	
AF group (Cardinality)	Features
Place (8)	Bilabial, alveolar, retroflex, palatal, velar, glottal, vowel, silence
Manner (6)	Plosive, fricative, approximant, nasal, vowel, silence
Roundness (4)	Rounded, unrounded, nil, silence
Frontness (5)	Front, mid, back, nil, silence
Height (6)	High, low, mid-high, mid-low, nil, silence

cardinality indicates the number of features in an AF group. Second column lists the possible feature values for each AF group. The articulatory feature specification for read speech is shown in Table 3.2. The AF specification in Table 3.2 differs by Table 5.1 in *place* AF group. This is because, the cardinality of *place* AF group of read speech is 9, whereas the cardinality of *place* AF group of extempore speech and conversation speech is 8. Higher cardinality of *place* AF group in read speech is due to the presence of labiodental feature value. The labiodental stands for sounds like /*v*/, but the Bengali speakers have a tendency to use /*bh*/ in place of /*v*/. Hence, the labiodental feature value is not found in *place* AF group of extempore and conversation modes of speech. However, we found very few instances of labiodental sound units in read speech, which is mainly because of the pronunciations of nouns involving /*v*/.

5.3.2 *Prediction of Articulatory Features*

The frame-level AFs for each AF group are predicted from spectral features using AF-predictors. The AF-predictors are developed using hidden Markov models (HMMs) and Feedforward neural networks (FFNNs). Separate AF-predictors are developed for each AF group of read, extempore, and conversation modes of speech. For training HMMs and FFNNs, to develop AF-predictors, the AF-level transcription is required. Hence, the phone labels are mapped to AF labels for extempore and conversation modes of speech.

Table 5.2 shows the mapping of each phone label into a set of AF labels of various AF groups for extempore and conversation modes of speech. First column lists the unique International Phonetic Alphabet (IPA) symbols found in IPA transcription. Second to sixth columns show the corresponding place, manner, roundness, frontness, and height AF values, respectively, for each phone. The mapping is derived using IPA chart [5]. The mapping of phone label to AF labels of different AF groups for read speech is shown in Table 3.3. The mapping of phone label to AF labels in Table 3.3 differs by Table 5.2 due to the presence of *labiodental* AF value in *place* AF group of Table 3.3. The *labiodental* AF value is absent in Table 5.2.

Table 5.2 Mapping of phone labels to AF groups for extempore and conversation modes of speech

Phones	Articulatory Feature Groups				
	Place	**Manner**	**Roundness**	**Frontness**	**Height**
a	vowel	vowel	unrounded	front	low
ɐ ɜ	vowel	vowel	unrounded	mid	mid-low
ɒ	vowel	vowel	rounded	back	low
ɑ	vowel	vowel	unrounded	back	low
æ ɛ	vowel	vowel	unrounded	front	mid-low
ə ɘ	vowel	vowel	unrounded	mid	mid-high
e	vowel	vowel	unrounded	front	mid-high
œ	vowel	vowel	rounded	front	mid-low
ɞ	vowel	vowel	rounded	mid	mid-low
i ɪ	vowel	vowel	unrounded	front	high
ʏ	vowel	vowel	rounded	front	high
ɔ	vowel	vowel	rounded	back	mid-low
o	vowel	vowel	rounded	back	mid-high
u ʊ	vowel	vowel	rounded	back	high
k k^h g g^h	velar	plosive	nil	nil	nil
ʧ $ʧ^h$ ʤ $ʤ^h$	palatal	plosive	nil	nil	nil
ʈ $ʈ^h$ ɖ $ɖ^h$	retroflex	plosive	nil	nil	nil
t t^h d d^h	alveolar	plosive	nil	nil	nil
p p^h b b^h	bilabial	plosive	nil	nil	nil
m	bilabial	nasal	nil	nil	nil
ɳ	retroflex	nasal	nil	nil	nil
ŋ	velar	nasal	nil	nil	nil
ɲ	palatal	nasal	nil	nil	nil
n	alveolar	nasal	nil	nil	nil
s ʃ ʒ ɵ ɬ	alveolar	fricative	nil	nil	nil
f v	bilabial	fricative	nil	nil	nil
h	glottal	fricative	nil	nil	nil
x	velar	fricative	nil	nil	nil
ʂ	retroflex	fricative	nil	nil	nil
j	palatal	approximant	nil	nil	nil
ɾ ɹ r l	alveolar	approximant	nil	nil	nil
ɭ	retroflex	approximant	nil	nil	nil
ʋ	bilabial	approximant	nil	nil	nil
sil	silence	silence	silence	silence	silence

The procedure for developing the AF-predictors using HMMs and FFNNs is described in Sect. 3.3.2.2. FFNNs with the size of input layer equal to 117 are used. The size of the hidden layer in the FFNNs is 585. The size of output layer for each AF group is equal to the cardinality of that AF group as shown in Table 5.1. Table 5.3

Table 5.3 Number of epochs carried out during training of FFNN-based AF-predictors for read, extempore, and conversation modes of speech

AF group	Number of epochs used for training		
	Read	Extempore	Conversation
Place	10	11	7
Manner	8	6	7
Roundness	8	6	6
Frontness	7	6	6
Height	9	10	8

shows the number of epochs carried out during training the FFNNs for various AF groups of read, extempore, and conversation modes of speech. First column indicates the AF group. Second, third, and fourth columns show the number of epochs carried out for read, extempore, and conversation modes of speech, respectively.

5.3.3 Performance Evaluation of AF-Predictors

The accuracy of AF-predictors is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 5.4 shows the accuracy of prediction of AFs for different AF groups of read, extempore, and conversation modes of speech. First column indicates the AF group. Second and third columns show AFs' prediction accuracies for read speech, while the fourth and fifth columns tabulate the AFs' prediction accuracies for extempore speech. Last two columns show the prediction accuracies for conversation speech. The results are shown separately for HMM-based and FFNN-based systems. It is observed that the prediction accuracy of all the AF groups is higher in FFNNs compared to HMMs for read and conversation modes of speech, while the prediction accuracy of most of the AF groups is higher in FFNNs compared to HMMs for extempore speech. Since FFNNs have higher recognition accuracies for all AF groups of read, conversation modes of speech and for majority of AF groups in extempore speech, we have used the FFNNs for predicting the AFs of various AF groups.

5.4 Articulatory Feature-Based Tandem Phone Recognition Systems

In this study, we have developed PRSs for read, extempore, and conversation modes of speech of Bengali using HMMs. The number of phones considered for developing PRSs for read, extempore, and conversation modes of speech is 35, 31, and 31, respectively. Most frequently occurring phones in the IPA transcription are considered for

Table 5.4 Prediction accuracy (%) of AF-predictors of different AF groups across read, extempore, and conversation modes of speech

AF group	Prediction accuracy (%) of AF-predictors					
	Read		Extempore		Conversation	
	HMMs	FFNNs	HMMs	FFNNs	HMMs	FFNNs
Place	55.04	70.35	51.26	62.39	48.72	61.97
Manner	67.51	74.40	63.57	68.19	56.25	65.65
Roundness	68.16	78.58	68.35	65.19	61.58	66.50
Frontness	67.64	74.01	64.37	60.99	58.66	66.48
Height	62.57	67.75	58.30	61.61	55.06	63.17

Table 5.5 Phone recognition accuracy (%) of AF-based tandem PRSs across read, extempore, and conversation modes of speech

Features	Recognition Accuracy (%)		
	Read	Extempore	Conversation
MFCCs (Baseline)	45.48	39.58	37.20
MFCCs + Place AFs	48.89	42.15	40.66
MFCCs + Manner AFs	47.74	41.11	40.18
MFCCs + Roundness AFs	47.28	40.46	38.45
MFCCs + Frontness AFs	46.59	40.75	38.85
MFCCs + Height AFs	48.60	42.93	39.40

building PRSs. HMM-based PRSs are developed using the procedure mentioned in section "Development of AF-Predictors using HMMs". The baseline PRSs are developed using MFCCs as features. AF-based tandem PRSs are developed using the combination of MFCCs and the predicted AFs as features. The AFs for each AF group are predicted from the spectral features using the FFNNs, as per the procedure mentioned in Sect. 3.3.2.2. Five AF-based tandem PRSs are developed separately, for read, extempore, and conversation modes of speech. Phone recognition accuracy is determined as per the procedure mentioned in Sect. 3.3.2.2. Table 5.5 shows the phone recognition accuracies of baseline and AF-based tandem PRSs of read, extempore, and conversation modes of speech. First column shows the different types of features used in the development of PRSs. Second, third, and fourth columns indicate the recognition accuracies obtained using read, extempore, and conversation modes of speech, respectively.

It is observed that all AF-based tandem PRSs have higher recognition accuracy compared to baseline PRSs in all three modes of speech. Among vowel AF groups, the *height* AF-based tandem PRSs have shown higher recognition accuracy in all the three modes of speech. Among consonant AF groups, the *place* AF-based tandem PRSs have shown higher recognition accuracy in all the three modes of speech. *Place* AF-based tandem PRSs of read and conversation modes of speech

have highest recognition accuracy, whereas the *height* AF-based tandem PRS has highest recognition accuracy in extempore mode of speech. The average improvement in the recognition accuracy of AF-based tandem PRSs using read, extempore, and conversation modes of speech is 2.34, 1.9, and 2.30%, respectively. The average improvement in the recognition accuracy is nearly same in read and conversation modes of speech, while the average improvement in the recognition accuracy is least in extempore speech compared other two modes.

5.5 Hybrid Phone Recognition Systems Using Articulatory Features

Hybrid PRSs are developed by combining AF-based tandem PRSs using weighted combination scheme. Three hybrid PRSs, namely (i) consonant-AF-based hybrid PRS, (ii) vowel-AF-based hybrid PRS, and (iii) all-AF-based hybrid PRS, are developed using the combinations mentioned in Sect. 3.5.1. PP-based tandem PRSs are developed across read, extempore, and conversation modes of speech to compare the performance of AF-based hybrid PRSs with PP-based tandem PRSs. Table 5.6 shows the optimal weighting factors used for developing hybrid PRSs of extempore and conversation modes of speech. The optimal weighting factors for developing hybrid PRSs of read speech are shown in Table 3.13. First column lists the different types of hybrid PRSs. Second to sixth columns indicate the weighting factors for extempore speech, while the last five columns indicate the weighting factors for conversation speech. The *hyphen* (-) symbol in Table 5.6 indicates that the particular weighting factor is not applicable for the corresponding hybrid PRS. The weighting factors w1, w2, w3, w4, and w5 correspond to place, manner, roundness, frontness, and height AF-based tandem PRSs, respectively. It can be observed that the AF-based tandem PRS having higher recognition accuracy in Table 5.5 will have a higher weighting factor in the corresponding AF group of Table 5.6 and vice versa. Further, we have also combined PP-based tandem PRSs and all-AF-based hybrid PRSs in all three modes of speech to develop PP-and-All-AF-based hybrid PRSs. The optimal weighting factors used for combining PP-based tandem PRSs and

Table 5.6 Weighting factors used for developing hybrid PRSs using weighted combination approach

Hybrid PRS	Weighting Factors									
	Extempore					Conversation				
	w1	w2	w3	w4	w5	w1	w2	w3	w4	w5
consonant-AF-based	0.6	0.4	–	–	–	0.5	0.5	–	–	–
vowel-AF-based	–	–	0.1	0.4	0.5	–	–	0.4	0.2	0.4
all-AF-based	0.3	0.2	0.1	0.1	0.3	0.4	0.1	0.1	0.1	0.3

Table 5.7 Phone recognition accuracy (%) of hybrid PRSs across read, extempore, and conversation modes of speech

PRSs using different features	Recognition accuracy (%)		
	Read	Extempore	Conversation
MFCCs (Baseline)	45.48	39.58	37.20
PP-based tandem PRS	48.97	40.60	42.14
consonant-AF-based Hybrid PRS	49.95	43.97	42.05
vowel-AF-based hybrid PRS	51.28	44.89	41.52
all-AF-based hybrid PRS	52.24	45.70	42.97
PP-and-All-AF-based hybrid PRS	52.61	46.24	44.15

all-AF-based hybrid PRSs are 0.2 and 0.8 for extempore speech and 0.4 and 0.6 for conversation speech. Since, in conversation speech, PP-based tandem PRS and all-AF-based hybrid PRS have almost the same performance, a nearly equal weightage is given to both PP-based tandem PRS and all-AF-based hybrid PRS by using the optimal weighting factors of 0.4 and 0.6. But, in extempore speech, PP-based tandem PRS has much lower performance than all-AF-based hybrid PRS; hence, a higher weightage is given to all-AF-based hybrid PRS than PP-based tandem PRS.

The performance of hybrid PRSs is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 5.7 shows the phone recognition accuracies of hybrid PRSs. First column lists different types of hybrid PRSs. Second, third, and fourth columns show the recognition accuracies of read, extempore, and conversation hybrid PRSs, respectively.

It is found that the performance of hybrid PRSs is higher than any of the AF-based tandem PRSs in all the three modes of speech. The improvement in the recognition accuracies of hybrid PRSs is consistent in all three modes of speech. Among consonant-AF-based and vowel-AF-based hybrid PRSs, the vowel-AF-based hybrid PRSs have higher recognition accuracy for read and extempore modes of speech, while the consonant-AF-based hybrid PRSs have higher recognition accuracy for conversation speech. It is observed that consonants have shown improvement in consonant-AF-based hybrid PRSs compared to vowels and it is vice versa for vowel-AF-based hybrid PRSs. In all-AF-based hybrid PRSs, both consonants and vowels have shown higher improvement compared to baseline PRSs. all-AF-based hybrid PRSs have higher recognition accuracy compared to PP-based tandem PRSs. The PP-and-All-AF-based hybrid PRSs have 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. Read speech has higher improvement in recognition accuracy compared to other two modes. The improvement in the performance of conversation speech is nearly same as that of extempore speech. The improvement in the recognition accuracy of read and extempore modes of speech is mainly due to the use of AFs, whereas much of the improvement for conversation speech is due to the use of PPs. Without the use of PPs, the highest improvement in the recognition accuracy of conversation speech obtained is 5.77%, which is less than that of extempore speech.

5.6 Phone Recognition Systems Using Excitation Source and Vocal Tract System Features

In this study, MFCC features are used for capturing the vocal tract information. The excitation source information from LP residual is parameterized into two feature sets, namely (i) RMFCCs and (ii) MPDSS. The procedure for developing HMM-based PRSs is mentioned in section "Development of AF-Predictors using HMMs". The number of phones considered for developing PRSs for read, extempore, and conversation modes of speech is 35, 31 and 31, respectively. Most frequently occurring phones in the transcription are considered for developing the PRSs. Phone recognition accuracy is determined as per the procedure mentioned in Sect. 3.3.2.3. Table 5.8 shows the recognition accuracy of PRSs developed using different types of features for read, extempore, and conversation modes of speech. First column shows the different types of features used in development of PRSs. Second, third, and fourth columns indicate the recognition accuracies of read, extempore, and conversation modes of speech, respectively.

From Table 5.8, it is observed that the use of excitation source information resulted in the improvement of phone recognition accuracy in all three modes of speech (see 6, 7, and 8 rows). The PRSs developed using excitation source features alone have poor recognition accuracy compared to the PRSs developed using MFCC features in all three modes of speech (see 2, 3, 4, and 5 rows). This indicates that discriminative ability of MFCCs to discriminate among various phones is higher compared to excitation source features. The phone recognition accuracy obtained using RMFCC features is higher than the phone recognition accuracy obtained using MPDSS features in all the three modes of speech. The combination of RMFCCs and MPDSS has shown higher recognition accuracy in all three modes of speech compared to either of RMFCCs or of MPDSS features alone. The combination of MFCCs, RMFCCs, and MPDSS has shown highest recognition accuracy in all three modes of speech. This shows that the combination of vocal tract and excitation source information helps

Table 5.8 Phone recognition accuracy (%) of PRSs developed using excitation source and vocal tract system features across read, extempore, and conversation modes of speech

Features	Recognition accuracy (%)		
	Read	Extempore	Conversation
MPDSS	11.40	11.73	8.28
RMFCC	25.72	24.28	21.68
RMFCC + MPDSS	27.30	26.73	22.33
MFCCs (Baseline)	45.48	39.58	37.20
MFCC + MPDSS	47.29	41.96	37.84
MFCC + RMFCC	48.31	42.78	38.16
MFCC + RMFCC + MPDSS	48.66	42.86	39.17

in better discrimination among different types of phones. The highest improvement obtained in the recognition accuracy of read, extempore, and conversation modes of speech is 3.18, 3.28, and 1.97%, respectively. Since the read and extempore modes of speech have similar characteristics, the improvement in the recognition accuracy is nearly same in read and extempore modes of speech.

It is observed that the improvement in the recognition accuracy of combination of MFCCs and excitation source features is mainly because of the improvement in classification accuracies of unaspirated stops. This is because, the excitation source features contain the information for discriminating between aspirated and unaspirated plosive consonants. The improvement in classification accuracy of unaspirated plosive consonants is mainly because of the reduction in the misclassification of unaspirated plosives to aspirated plosives. Since the conversation speech has less number of unaspirated consonants compared other two modes, the improvement obtained due to the reduction of misclassification of unaspirated plosives to aspirated plosives is less. Hence, the overall improvement in the recognition accuracy of conversation speech is less compared to read and extempore modes of speech.

5.7 Analysis Across Read, Extempore, and Conversation Modes of Speech

The performance of PRSs is analyzed in all three modes of speech by considering five broad phonetic subgroups. The five broad phonetic subgroups considered are plosives, nasals, fricatives, vowels, and approximants (semivowels). The performance of PRSs using five subgroups is shown in Table 5.9. First column shows the features used for developing PRSs. Second, third, and fourth columns show the recognition accuracies obtained using read, extempore, and conversation modes of speech, respectively. From Table 5.9, it can be found that the performance of PRSs using either articulatory or excitation source features in addition to vocal tract features is higher than the performance of PRSs using vocal tract information alone. The combination of MFCCs and AFs has shown higher recognition accuracy compared to the combination of MFCCs and excitation source features. The highest improvement obtained at subgroup level for read, extempore, and conversation modes of speech is

Table 5.9 Phone recognition accuracy (%) of PRSs by considering five broad phonetic subgroups

Features	Recognition accuracy (%)		
	Read	Extempore	Conversation
MFCCs	75.69	66.88	66.55
MFCCs + RMFCCs + MPDSS	76.75	69.76	67.85
PP-and-All-AF-based hybrid PRS	79.04	71.35	69.04

3.35, 4.47, and 2.49%, respectively. Although the PRSs based on five broad phonetic subgroups cannot be directly used for developing speech recognition systems, they can be used as a first-level classifiers in some applications such as automatic language identification systems and audio retrieval systems. It can be observed that the extempore and conversation PRSs, which had less than 40% phone-level accuracies with MFCCs, have around 70% subgroup level accuracies with MFCCs and AFs. Hence, the use of articulatory and excitation source features is very effective across all three modes of speech.

The analysis of PRSs developed using five broad phonetic subgroups in three modes of speech is as follows: In read speech, it is observed that the vowels, approximants, and plosives are more accurately detected using AFs than excitation source features, while the nasals and fricatives have better classification accuracy in PRSs using excitation source features than that of AFs. The detection of silence is more accurate in both AFs and excitation source features compared to spectral features. We can exploit the advantages of both AF-based system and excitation source feature-based system by using the AF-based system to recognize vowels, approximants, and plosives and the excitation source feature-based system to recognize nasals and fricatives. In extempore mode of speech, it is observed that the fricatives and nasals perform better with AFs, while the plosives and vowels have higher classification accuracy using excitation source features. We can take benefit from both the systems, by using AFs to recognize nasals and fricatives and excitation source features to recognize vowels and plosives. This kind of combination will lead to much better improvement at subgroup level. Semivowels have lowest classification accuracy. The misclassification mainly exists between vowels and semivowels. However, the misclassification of vowels into semivowels has reduced in both AF-based and excitation source feature-based systems. In conversation mode of speech, fricatives have higher classification accuracy with both AFs and excitation source features compared to MFCCs. Nasals and vowels are more accurately recognized in AF-based systems. The plosives have higher classification accuracy with the excitation source features compared to AFs. We can further improve the overall performance of system by combining in such a way that the fricatives and plosives are recognized using excitation source features, while the nasals and vowels are recognized using AFs, and the approximants are decoded using spectral features. It is found that plosives are mainly misclassified to approximants, because of the confusion between voiced plosives and approximants. Nasals have least classification accuracy, which are mainly misclassified into approximants. This is because, both nasals and approximants are sonorants and both have similar characteristics. In general, it is observed that the excitation source features have higher recognition accuracy for plosives and fricatives in all three modes of speech, whereas the nasals and vowels have better recognition accuracy using AFs. Generally, the approximants have higher classification accuracy with spectral features.

The reasons for misclassification of phones are examined across read, extempore, and conversation modes of speech. In case of read speech, the sentences which are read very fast have more number of errors. This is because, locating the phones in the speech signal, even manually, is very difficult in the sentences which are read

very fast, i.e., all the perceived sound units are not present in the speech signal. The more number of errors in extempore speech is due to the presence of long pauses (silences). In extempore speech, speakers have a tendency to leave long pauses, while thinking for what needs to be delivered next. The long pauses (silences) are misclassified into unvoiced consonants. The errors in conversation speech are due to the following reasons: (i) Speakers have a tendency to use certain words of other language such as English, while having a conversation in Bengali, (ii) speakers speak very fast in conversation such that all the perceived sound units can not be located in the speech signal, and (iii) presence of background noises or the noises introduced by the communication channels, in case of the conversation data collected from television or radio channels.

We have also analyzed the recognition errors with respect to position of the sound units in all three modes of speech. In read speech, in case of two consecutive vowels or consecutive vowel semivowel pair, only one vowel is recognized. The word $\{inouka\}$ is recognized as $\{inuka\}$, where $\{ou\}$ is decoded as $\{u\}$. If a consonant is repeated twice in a word, then it is recognized as single consonant. The word $\{jammu\}$ is decoded as $\{jamu\}$. If there are two consecutive words such that the ending of the first word and the beginning of the second word both are unvoiced consonants like $\{k, p, t\}$, then one of unvoiced consonant is omitted by the recognizer. The pair of two consecutive words $\{kishap\ kode\}$ is recognized as $\{kishap\ ode\}$, where $\{k\}$ present in the beginning of the second word is missed. In extempore speech, the problems due to repeated consonants and consecutive vowels as explained in read speech are observed. Along with them, there are few other problems which are listed as follows. If the word is spoken very fast, then some of phones in the middle of the word will not be recognized. If a word starts after a silence and the beginning of the word is an unvoiced consonant, such as $\{k, p, t\}$, then the unvoiced consonants in the beginning of the word will not be recognized. For example, $\{< silence > kibhabe\}$ will be recognized as $\{< silence > ibhabe\}$. In the words ending with a Consonant-Vowel-Consonant (CVC) syllable, the last *consonant* will be omitted by the recognizer. The word $\{kabor\}$ will be decoded as $\{kabo\}$, where $\{r\}$ present at the end of the word is missed. All the errors which occur in both read and extempore modes of speech are also observed in conversation mode of speech. But, the errors due to CVC syllable present in the end of a word are very severe. Since the conversation speech is generally spoken very fast, there are lot of errors in the middle of the words. The length (duration) of the phones present in the middle of the words is extremely less. Many phones present in the middle of the word are not recognized, which is a major source of error in conversation mode of speech.

Further, the performance of PRSs is analyzed in all three modes of speech by merging all the unaspirated consonants to aspirated consonants. Table 5.10 shows the improvement in the recognition accuracy of PRSs in three modes of speech after merging unaspirated consonants to aspirated consonants. The improvement in the baseline PRSs is compared with the PRSs developed using the combination of MFCCs and excitation source features.

From Table 5.10, it can be found that the improvement in the performance, after merging the unaspirated consonants to aspirated consonants, is higher in baseline

Table 5.10 Improvement (%) in the performance of PRSs across read, extempore, and conversation modes of speech after merging unaspirated and aspirated consonants

Features	Recognition Accuracy (%)		
	Read	Extempore	Conversation
MFCCs (Baseline)	1.62	1.59	2.24
MFCC + RMFCC + MPDSS	1.22	0.49	1.63

PRSs compared to the PRSs using the combination of MFCCs and excitations source features in all three modes of speech. The improvement obtained by merging aspirated and unaspirated consonants is less in case of the PRSs developed using combination of spectral and excitation source features. This is because, the excitation source features used for developing the PRSs have reduced the misclassification between unaspirated and aspirated consonants. Hence, it is clear that the use of excitation source features results in reduction of misclassification between unaspirated and aspirated consonants. It is also observed that the use of excitation source features results in reduction of misclassification among the pairs of phones with same manner and place of articulation, but differs only in their excitation in all the three modes of speech. This clearly indicates that the use of excitation source features is responsible for improving the recognition accuracy in all the three modes of speech.

The reasons for higher recognition accuracy of read speech compared to extempore and conversation modes of speech are as follows: Read speech involves reading out from the notes and uses a more formal language. The amount of phonetic and prosodic information captured in the read speech is more stable and systematic, compared to extempore and conversation modes of speech. Since the read speech is prepared well in advance and delivered in a more structured and constrained way, the quality of read speech is much better compared to extempore and conversation modes of speech. In case of read speech, almost all the perceived sound units could be located in the speech signal.

Extempore speech is delivered spontaneously without the aid of notes. Hence, it has several irregularities, such as uneven (non-uniform) pauses and unexpected breaks. These irregularities result in poor phonetic and unstructured prosodic information.

In case of conversation speech, most of the sentences are spoken very fast and locating the phones in the speech signal, even manually, is very difficult. All the perceived sound units could not be located in the speech signal. The speakers have a tendency to use certain words of other language such as English, while having a conversation in Bengali, which leads to more number of errors. In case of the conversation data, which is collected from television or radio channels, there exists background noises or the noises introduced by the communication channels, and it results in poor quality of the speech signal.

Hence, the overall quality of read speech is better than conversation and extempore modes of speech. The characteristics of most of the sound units in read speech are steady and stable, whereas in case of extempore and conversation modes of speech, the characteristics of sound units are not stable and lot of variance is observed. Hence,

in our studies, we have observed better accuracy in case of read speech compared to extempore and conversation modes of speech. Since the quality of extempore speech is better than conversation speech, the recognition accuracy of extempore speech is better than that of conversation speech.

5.8 Summary

In this chapter, the performance of PRSs across read, extempore, and conversation modes of speech is analyzed using articulatory and excitation source features. The combination of articulatory and spectral features has led to the improvement of recognition accuracy in all three modes of speech. Hybrid PRSs are developed and compared across read, extempore, and conversation modes of speech. all-AF-based hybrid PRSs outperform the conventional PP-based tandem PRSs in all three modes of speech. PP-and-All-AF-based hybrid PRSs have 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. Read speech has higher improvement in recognition accuracy compared to other two modes. The improvement in the performance of conversation speech is nearly same as that of extempore speech. The improvement in the recognition accuracy of read and extempore modes of speech is mainly due to the use of AFs, whereas much of the improvement for conversation speech is due to the use of PPs. The use of excitation source information in addition to vocal tract information has improved the performance of PRSs across 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. The use of excitation source features for developing PRSs reduces the misclassification between unaspirated and aspirated plosives, which leads to the improvement of phone recognition accuracy. Among the three 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. The improvement obtained in the performance using AFs is much higher compared to the improvement obtained using excitation source features.

References

1. 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
2. Manjunath K.E., K. Sreenivasa Rao, Articulatory and excitation source features for speech recognition in read, extempore and conversation modes. Int. J. Speech Technol. (2015), pp. 1–14

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. The International Phonetic Association, *Handbook of the International Phonetic Association*, Cambridge University Press, http://www.langsci.ucl.ac.uk/ipa/index.html

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)

Appendix A
MFCC Features

The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. The detailed description of various steps involved in the MFCC feature extraction is explained below.

1. **Pre-emphasis**: Pre-emphasis refers to filtering that emphasizes the higher frequencies. Its purpose is to balance the spectrum of voiced sounds that have a steep roll-off in the high-frequency region. For voiced sounds, the glottal source has an approximately -12 dB/octave slope [1]. However, when the acoustic energy radiates from the lips, this causes a roughly $+6$ dB/octave boost to the spectrum. As a result, a speech signal when recorded with a microphone from a distance has approximately a -6 dB/octave slope downward compared to the true spectrum of the vocal tract. Therefore, pre-emphasis removes some of the glottal effects from the vocal tract parameters. The most commonly used pre-emphasis filter is given by the following transfer function

$$H(z) = 1 - bz^{-1} \tag{A.1}$$

   where the value of b controls the slope of the filter and is usually between 0.4 and 1.0 [1].
2. **Frame blocking and windowing**: The speech signal is a slowly time-varying or quasi-stationary signal. For stable acoustic characteristics, speech needs to be examined over a sufficiently short period of time. Therefore, speech analysis must always be carried out on short segments across which the speech signal is assumed to be stationary. Short-term spectral measurements are typically carried out over 20 ms windows, and advanced every 10 ms [2, 3]. Advancing the time window every 10 ms enables the temporal characteristics of individual speech sounds to be tracked, and the 20 ms analysis window is usually sufficient to provide good spectral resolution of these sounds, and at the same time short enough to resolve significant temporal characteristics. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered

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

at some frame. On each frame, a window is applied to taper the signal towards the frame boundaries. Generally, Hanning or Hamming windows are used [1]. This is done to enhance the harmonics, smooth the edges, and to reduce the edge effect while taking the DFT on the signal.

3. **DFT spectrum**: Each windowed frame is converted into magnitude spectrum by applying DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; \qquad 0 \leq k \leq N-1 \tag{A.2}$$

where N is the number of points used to compute the DFT.

4. **Mel spectrum**: Mel spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as Mel-filter bank. A Mel is a unit of measure based on the human ears perceived frequency. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly. The Mel scale is approximately a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz [4]. The approximation of Mel from physical frequency can be expressed as

$$f_{Mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{A.3}$$

where f denotes the physical frequency in Hz, and f_{Mel} denotes the perceived frequency [2].

Filter banks can be implemented in both time domain and frequency domain. For MFCC computation, filter banks are generally implemented in frequency domain. The center frequencies of the filters are normally evenly spaced on the frequency axis. However, in order to mimic the human ears perception, the warped axis, according to the nonlinear function given in Eq. (A.3), is implemented. The most commonly used filter shaper is triangular, and in some cases the Hanning filter can be found [1]. The triangular filter banks with Mel frequency warping is given in Fig. A.1.

The Mel spectrum of the magnitude spectrum $X(k)$ is computed by multiplying the magnitude spectrum by each of the of the triangular Mel weighting filters.

$$s(m) = \sum_{k=0}^{N-1} \left[|X(k)|^2 H_m(k) \right]; \qquad 0 \leq m \leq M-1 \tag{A.4}$$

where M is total number of triangular Mel weighting filters [5, 6]. $H_m(k)$ is the weight given to the k^{th} energy spectrum bin contributing to the m^{th} output band and is expressed as:

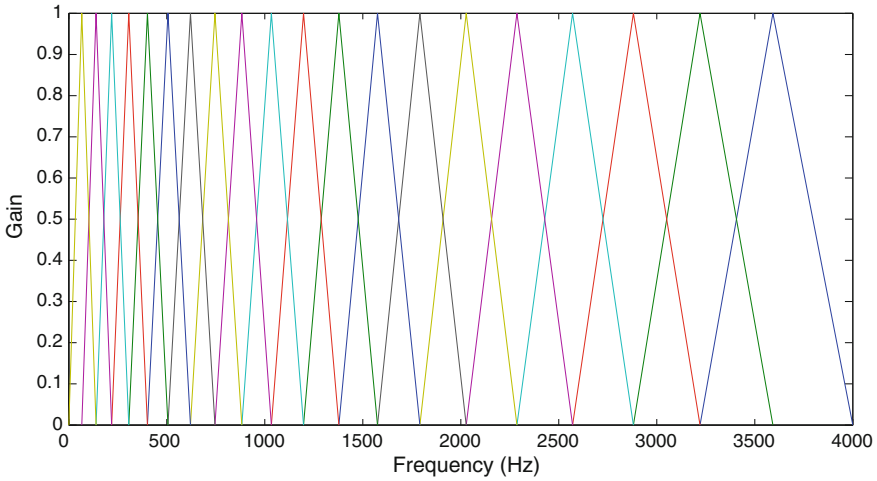

Fig. A.1 Mel-filter bank

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{A.5}$$

with m ranging from 0 to $M-1$.

5. **Discrete cosine transform (DCT)**: Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed Mel frequency coefficients produces a set of cepstral coefficients. Prior to computing DCT, the Mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a quefrequency peak corresponding to the pitch of the signal and a number of formants representing low quefrequency peaks. Since most of the signal information is represented by the first few MFCC coefficients, the system can be made robust by extracting only those coefficients ignoring or truncating higher order DCT components [1]. Finally, MFCC is calculated as [1]

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(s(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \qquad n = 0, 1, 2, \ldots, C-1 \tag{A.6}$$

where $c(n)$ are the cepstral coefficients, and C is the number of MFCCs. Traditional MFCC systems use only 8–13 cepstral coefficients. The zeroth coefficient is often excluded since it represents the average log-energy of the input signal, which only carries little speaker-specific information.

6. **Dynamic MFCC features**: The cepstral coefficients are usually referred to as static features, since they only contain information from a given frame. The extra information about the temporal dynamics of the signal is obtained by computing first and second derivatives of cepstral coefficients [7–9]. The first-order derivative is called delta coefficients, and the second-order derivative is called delta–delta coefficients. Delta coefficients tell about the speech rate, and delta–delta coefficients provide information similar to acceleration of speech. The commonly used definition for computing dynamic parameter is [7]

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \tag{A.7}$$

where $c_m(n)$ denotes the m^{th} feature for the n^{th} time frame, k_i is the i^{th} weight, and T is the number of successive frames used for computation. Generally T is taken as 2. The delta–delta coefficients are computed by taking the first-order derivative of the delta coefficients.

References

1. J.W. Picone, Signal modeling techniques in speech recognition. Proc. IEEE **81**, 1215–1247 (1993)
2. J.R. Deller, J.H. Hansen, J.G. Proakis, *Discrete Time Processing of Speech Signals* (Prentice Hall, NJ, 1993)
3. J. Benesty, M.M. Sondhi, Y.A. Huang, *Handbook of Speech Processing* (Springer, New York, 2008)
4. J. Volkmann, S. Stevens, E. Newman, A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. **8**, 185–190 (1937)
5. Z. Fang, Z. Guoliang, S. Zhanjiang, Comparison of different implementations of MFCC. J. Comput. Sci. Technol. **16**, 582–589 (2000)
6. G.K.T. Ganchev, N. Fakotakis, Comparative evaluation of various MFCC implementations on the speaker verification task, in *Proceedings of International Conference on Speech and Computer (SPECOM)* (2005), pp. 191–194
7. L. Rabiner, B.-H. Juang, B. Yegnanarayana, *Fundamentals of Speech Recognition* (Pearson Education, London, 2008)
8. S. Furui, Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Trans. Acoust. Speech Sig. Proc. **29**, 342–350 (1981)
9. J.S. Mason, X. Zhang, Velocity and acceleration features in speaker recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1991), pp. 3673–3676

Appendix B
Pattern Recognition Models

In this work, hidden Markov model (HMM), support vector machine (SVM), and auto-associative neural network (AANN) models are used to capture the pattern present in features. HMMs are used to capture the sequential information present in feature vectors for CV recognition. SVMs are used to capture the discriminative information present in the feature vectors for CV recognition. AANN models are used to capture the nonlinear relations among the feature vectors for speaker identification. The following sections briefly describe the pattern recognition models used in this study.

B.1 Hidden Markov Models

Hidden Markov models (HMMs) are the commonly used classification models in speech recognition [1]. HMMs are used to capture the sequential information present in feature vectors for developing PRSs. HMM is a stochastic signal model which is referred to as Markov sources or probabilistic functions of Markov chains. This model is an extension to the concept of Markov model which includes the case where the observation is a probabilistic function of the state. HMM is a finite set of states, each of which is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state, an outcome or observation can be generated, according to the associated probability distribution. Here, only the outcome is known and the underlying state sequence is hidden. Hence, it is called a hidden Markov model.

Following are the basic elements that define HMM:

1. N, Number of states in the model,
   $s = \{s_1, s_2, ...s_N\}$
2. M, Number of distinct observation symbol per state,
   $v = \{v_1, v_2, ...v_M\}$
3. State transition probability distribution $A = \{a_{ij}\}$, where

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

$$a_{ij} = P\left[q_{t+1} = s_j | q_t = s_i\right], 1 \leq i, j \leq N \tag{B.1}$$

4. Observation symbol probability distribution in state j,
   B = $\{b_j(k)\}$, where

$$b_j(k) = P\left[v_k \; at \; t | q_t = s_j\right] \qquad 1 \leq j \leq N, 1 \leq k \leq M \tag{B.2}$$

5. Initial state distribution $\Pi = \{\Pi_j\}$, where

$$\Pi_j = P\left[q_1 = s_i\right] \qquad 1 \leq i \leq N \tag{B.3}$$

So, a complete specification of an HMM requires specification of two model parameters (N and M), specification of observation symbols, and the specification of three probability measures A, B, Π. Therefore, HMM is indicated by the compact notation

$$\lambda = (A, B, \Pi)$$

Given that state sequence $q = (q1q2...q_T)$ is unknown, the probability of observation sequence $O = (o1o2...o_T)$, given the model λ, is obtained by summing the probability of over all possible state sequences q as follows:

$$P(o|\lambda) = \sum_{q_1, q_2, \ldots, q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) ... a_{q_{T-1} q_T} b_{q_T}(o_T) \tag{B.4}$$

where π_{q_1} is the initial state probability of q_1, and T is length of observation sequence.

B.2 FeedForward Neural Networks

FFNNs are the artificial neural networks, where the information moves from the input layer to output layer through the hidden layer in forward direction with no loops in the network. FFNNs are used to capture the nonlinear relationship between the feature vectors and the phonetic sound units. FFNNs map an input feature vector into one of the phonetic units, among the set of phonetic sound units used for training the FFNN models. Each unit in one layer of the FFNN has directed connections to the units in the subsequent layer. FFNNs consist of an input layer, an output layer, and one or more hidden layers. The number of units in the input is equal to the dimension of feature vectors, while the number of units in output layer is equal to the number of phonetic sound units being modeled. The hidden and output layers are nonlinear, whereas the input layer is linear. The nonlinearity is achieved using activation functions such as sigmoid, softmax. The general structure of three-layered FFNN is as shown in Fig. B.1. A three-layered FFNN has one input layer, one hidden layer, and one output layer.

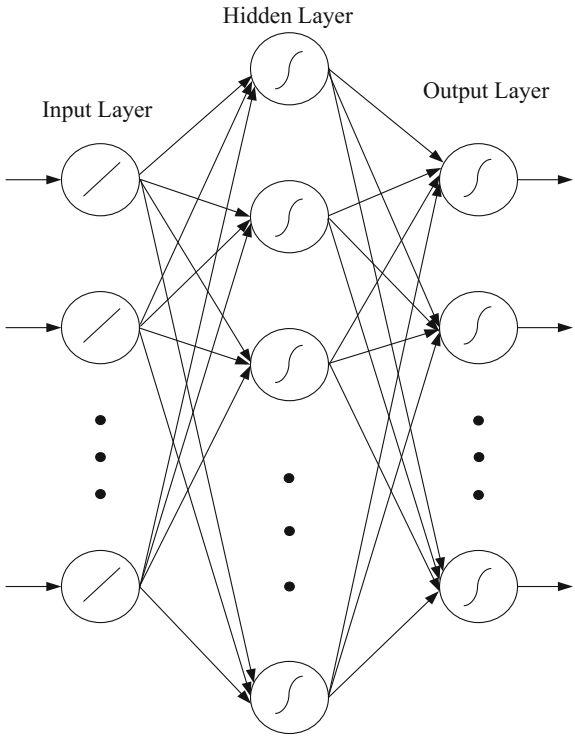

Fig. B.1 General structure of three-layered FeedForward neural networks

The feature vectors are fed to the input layer, and the corresponding phone labels are fed to the output layer of the FFNN. FFNNs are trained using a learning algorithm such as back-propagation algorithm [2, 3]. The back-propagation algorithm is most commonly used in the development of speech recognition applications using FFNNs. In back-propagation algorithm, the calculated output is compared with the correct output, and the error between them is computed using a predefined error function. The error is then back-propagated through the network, and the weights of the network are adjusted based on the computed error. The weights are adjusted using a nonlinear optimization method such as gradient descent method. This process is repeated for sufficiently large number of training examples till the network converges. After the completion of training phase, the weights of the network are used for decoding the phonetic sound units in the spoken utterances. Determining the network structure is an optimization problem. At present, there are no formal methods for determining the optimal structure of a neural network. The key factors that influence the neural network structure are amount of training data, learning ability of the network, and capacity to generalize the acquired knowledge.

References

1. L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989)
2. R. Rojas, *Neural Networks - A Systematic Introduction* (Springer, Berlin, 1996)
3. M. Nielsen, Neural Networks and Deep Learning. http://neuralnetworksanddeeplearning.com.