Articulatory and excitation source features for speech recognition in read, extempore and conversation modes

K. E. Manjunath[1] · K. Sreenivasa Rao[1]

Received: 21 August 2015 / Accepted: 28 November 2015 / Published online: 11 December 2015

Abstract In our previous works, we have explored articulatory and excitation source features to improve the performance of phone recognition systems (PRSs) using read speech corpora. In this work, we have extended the use of articulatory and excitation source features for developing PRSs of extempore and conversation modes of speech, in addition to the read speech. It is well known that the overall performance of speech recognition system heavily depends on accuracy of phone recognition. Therefore, the objective of this paper is to enhance the accuracy of phone recognition systems using articulatory and excitation source features in addition to conventional spectral features. The articulatory features (AFs) are derived from the spectral features using feedforward neural networks (FFNNs). 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 Mel frequency cepstral coefficients (MFCCs) and AFs derived from FFNNs. Hybrid PRSs are developed by combining the evidences from AF-based tandem PRSs using weighted combination approach. The excitation source information is derived by processing the linear prediction residual of the speech signal. The vocal tract information is captured using MFCCs. The combination of vocal tract and excitation source features is used for developing PRSs. The PRSs are developed using hidden Markov models. Bengali speech database is used for developing PRSs of read, extempore and conversation modes of speech. The results are analyzed and the performance is compared across different modes of speech. From the results, it is observed that the use of either articulatory or excitation source features along-with to MFCCs will improve the performance of PRSs in all three modes of speech. The improvement in the performance using AFs is much higher compared to the improvement obtained using excitation source features.

Keywords Read · Extempore and conversation modes of speech · Articulatory features · Linear prediction (LP) residual · Excitation source features · Phone recognition · Tandem phone recognition systems · HMMs · FFNNs

1 Introduction

Speech is produced by the air exhaled from the lungs which leads to the vibration of vocal folds. Further, the air passes through the vocal tract and then radiates out through nostrils or lips. The articulators such as lips, teeth, tongue, alveolar ridge, hard palate, velum and glottis are involved in speech production. The positioning and movement of various articulators during the production of a sound unit is represented using articulatory features (AFs). The AFs change from one sound unit to another (Gerfen 2015). Detailed description of AFs is given in Manjunath and Sreenivasa Rao (2015b), Manjunath et al. (2015a). From the theory of source-filter model of speech production, speech is produced by exciting a linear acoustic filter with an excitation source. The vocal-folds form the main source of excitation and the vocal tract can be viewed as linear acoustic filter. The variations in the vocal tract shape can be captured using time varying filter in the form of resonances and

✉ K. Sreenivasa Rao
ksrao@sit.iitkgp.ernet.in; ksrao@iitkgp.ac.in

K. E. Manjunath
ke.manjunath@gmail.com

[1] School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

antiresonances of speech spectrum. Just a mere shape of vocal tract without an excitation source would not be sufficient to produce speech. There are many consonants like $\{p, b\}$, which are produced due to same vocal tract shape, but differ in type of excitation. Detailed description of excitation source features is given in Manjunath and Sreenivasa Rao (2015a), Manjunath et al. (2015b). Hence, the production of speech is characterized by articulatory and excitation source features in addition to vocal tract features. The use of articulatory and excitation source features alongwith spectral features helps in better discrimination among different classes of phones. But, generally the phone recognition systems (PRSs) are developed using spectral features alone. The parameterization techniques such as linear prediction cepstral coefficients (LPCCs) and Mel-frequency cepstral coefficients (MFCCs) are used to capture spectral features representing the vocal tract information. Hence, there is a need for investigating articulatory and excitation source features in addition to spectral features for developing PRSs.

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 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.
- *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 a 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.

In order to perform phone recognition across three modes of speech in an efficient way, it is required to develop separate PRSs for each mode of speech. AFs contain lexical and phonetic information. The speech variability such as co-articulation 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. Similarly, the use of excitation source features improve the discrimination among different phonetic units. Hence, the objective of this work is to explore articulatory and excitation source features across three modes of speech with an intent to improve phone recognition accuracy.

The rest of the work is organized as follows: Sect. 2 provides the literature survey of articulatory and excitation source features. Section 3 describes the speech corpus used in present work. Section 4 discusses different types of feature extraction techniques employed. Section 5 describes the development of PRSs using articulatory and spectral features. The development of PRSs using vocal tract and excitation source features is described in Sect. 6. Section 7 provides the analysis of PRSs across three modes of speech. The summary and conclusion of the paper is presented in Sect. 8.

2 Literature survey

The area of speech recognition is one of the most active area of research from last six decades. The most common approaches of developing speech recognition systems use hidden Markov models (HMMs) (Lee and Hon 1989), feedforward neural networks (FFNNs) (Fallside et al. 1990) and combination of HMMs and FFNNs (Bourlard and Morgan 1994). The tandem (Hermansky et al. 2000; Ketabdar and Bourlard 2008) speech recognition systems are most commonly used to improve the performance of speech recognition systems. In recent years, deep learning is widely used to gain dramatic improvements in the performance of speech recognition systems. Mohamed et al. (2012), Hinton et al. (2012) have used deep neural networks for speech recognition. Graves et al. (2013) have explored deep recurrent neural networks for speech recognition. Sainath et al. (2013) explored convolutional neural networks for large vocabulary speech recognition. Toth (2014) proposed the use of maxout activation function for convolutional neural networks (CNNs) to improve performance of CNN-based speech recognition systems. In all of the above mentioned state-of-the-art systems the acoustic input is represented using MFCCs, which mainly capture vocal tract information. But, we can improve the recognition accuracy of PRSs using the articulatory and excitation source features in addition to vocal tract features.

There are some studies exploring articulatory and excitation source features to improve the phone recognition accuracy. Few of them are briefly discussed below. He et al. (1996) have used linear prediction (LP) residual features to improve the performance of isolated-word recognizer. Chengalvarayan (1998) has used LP residual features to improve the performance of city-name recognizer. Kirchhoff et al. (2002) have used AFs to develop robust speech recognition systems. Metze (2005) has used the AFs to improve the performance of conversational speech recognition systems. Siniscalchi and Lee (2009) have used the place and manner of articulation to improve the performance of the speech recognition systems. Dhananjaya et al. (2011) have used excitation source information to classify the manner of articulation accurately. Mitra et al. (2013) have used the articulatory trajectories information to improve the recognition accuracy of clean and noisy speech data. Manjunath et al. (2013), Manjunath et al. (2013), phone and consonant-vowel recognition systems were developed for Indian languages Odia and Bengali.

In all of the existing works, the articulatory and excitation source features are mainly explored using read speech corpora. There are few works exploring AFs for conversation speech, while there are no works exploring either articulatory and excitation source features for extempore mode of speech. Hence, in this work, we have proposed articulatory and excitation source features to develop PRSs in read, extempore and conversation modes of speech. It is found that, there are very limited number of works exploring articulatory and excitation source features in the context of Indian languages. Hence, in this work, we have used an Indian language Bengali to study articulatory and excitation source features across three modes of speech.

3 Speech corpus

For developing and analyzing the performance of the proposed phone recognition systems, speech corpora of Bengali language is considered. The phonetic and prosodically rich transcribed (PPRT) Bengali speech corpus developed at IIT Kharagpur is used in this study (Sunil Kumar et al. 2013). The speech corpus contains speech data collected in read, extempore and conversation modes of speech. The duration of read speech is 1.16 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 is transcribed using International Phonetic Alphabet (IPA) chart. IPA provides one symbol for each distinctive sound. IPA contains unique symbols for denoting 59 consonants, 35 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 (The International Phonetic Association 2015). 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 1 shows the number of speakers and the number of sentences used in this study. The details 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.

Table 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

4 Feature extraction

In this section, the feature extraction techniques to derive the articulatory and excitation source features are discussed. Spectral features are represented using MFCCs. The excitation source features are obtained by processing the LP residual of the speech signal, while the AFs are derived from the spectral features using pattern recognition models. The MFCCs and excitation source features are extracted using the procedure mentioned in Manjunath and Sreenivasa Rao (2015a). The detailed description for extracting the AFs is described in the following subsections.

4.1 Extraction of articulatory features

Articulatory features 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. 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. Table 2 shows the articulatory feature specification for read, extempore and conversation modes of speech. The AF specification of extempore and conversation modes of speech differs by that of read speech in *place* AF group. This is because, the cardinality of *place* AF group of read speech is 9, where as the cardinality of *place* AF group of extempore 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* /.

4.1.1 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 1 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 (Manjunath and Sreenivasa Rao 2015b; Manjunath et al. 2015a).

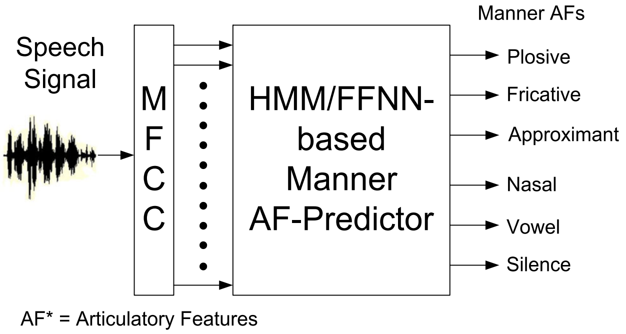

Fig. 1 Block diagram of prediction of manner articulatory features

Similar kinds of AF-predictors are developed for all five AF groups, as shown in Fig. 2. AFs for a particular AF group are predicted using the AF-predictor of that specific group.

4.1.2 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 2.

Table 3 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 (The International Phonetic

Table 2 Articulatory feature specification for read, extempore and conversation modes of speech

AF group (cardinality)	Features
Read speech	
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
Extempore and conversation modes of speech	
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

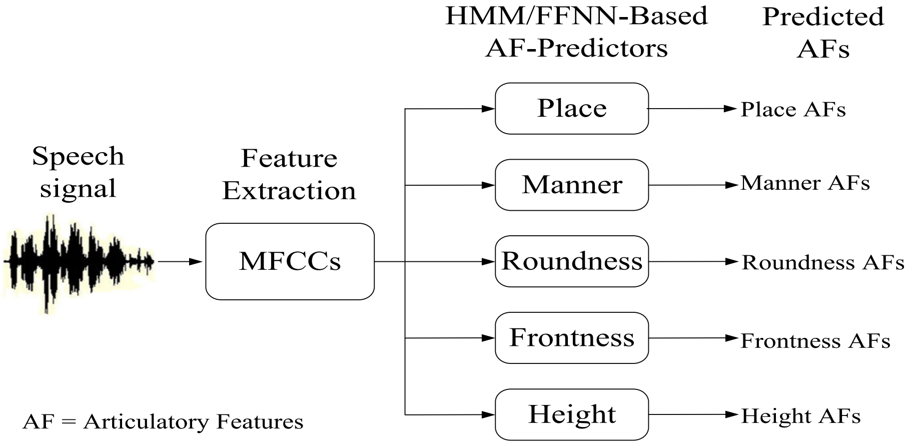

Fig. 2 Block diagram of the prediction of articulatory features

Association 2015). Similar kind of mapping for read speech is defined in Manjunath and Sreenivasa Rao (2015b). The mapping of phone label to AF labels of extempore and conversation modes differ by that of read speech due to presence of *labiodental* AF value in *place* AF group of read speech.

4.1.3 Development of AF-predictors using HMMs

In this study, HMM-based AF-predictors 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 AF group. HMMs are trained using maximum likelihood approach. The global *means* and *variances* are are used to create flat-start HMMs. The embedded reestimation using Baum-Welch algorithm followed by the Viterbi decoding is used. The open source HTK toolkit is used for building HMM models (The Hidden Markov Model Toolkit and HTK book 2015).

4.1.4 Development of AF-predictors using FFNNs

The procedure for developing FFNN-based AF-predictors 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 labels 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 perhaps 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 non-linearity activation function is used at output layer to constrain posterior probabilities to lie between zero and one and sum to one. The weights associated to 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 (Speech Group at the International Computer Science Ins 2010).

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 three frames with a duration of 45 ms. The number of nodes in input layer (NNIL) is determined using Eq. 1.

$$NNIL = No.\,of\,frames\,in\,temporal\,context \times Dimension\,of\,MFCCs \qquad (1)$$

Table 3 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 3	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
Y	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 ʤʤ^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
s	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

According to Eq. 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. Table 4 shows the number of epochs carried out during training the FFNNs for various AF groups of read, extempore and conversation modes of speech.

4.1.5 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 (The Hidden Markov Model Toolkit and HTK book 2015). Once the optimal alignment is found, the number of substitution errors (S), deletion errors (D) and insertion errors (I) are determined and the percentage accuracy is calculated using Eq. 2.

Table 4 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

$$Percentage\,Accuracy = \frac{N - D - S - I}{N} \times 100\,\% \tag{2}$$

where N is the total number of labels in the reference transcriptions. Table 5 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 tabulates 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 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.

4.2 Prediction of phone posterior features

Phone posteriors (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

$$\sum_{i=1}^{N} P(i) = 1, \tag{3}$$

where N—total number of phone classes; i—indicates specific phone class.

FFNN is trained, for predicting the PPs, using the procedure mentioned in Sect. 4.1.4. The weights associated to 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 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-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 (Ketabdar and Bourlard 2008). The dimension of generated PPs will be equal to the number of phones considered for training FFNNs. We have used a temporal context of three 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 (Ketabdar and Bourlard 2008).

Table 5 Prediction accuracy (%) of AF-predictors of different AF groups across read, extempore and converstaion 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

5 Phone recognition systems using articulatory and spectral features

This section discusses the development of PRSs using the combination of articulatory and spectral features. Tandem PRSs are developed using AFs. The tandem PRSs are then combined using weighted combination scheme to develop hybrid PRSs.

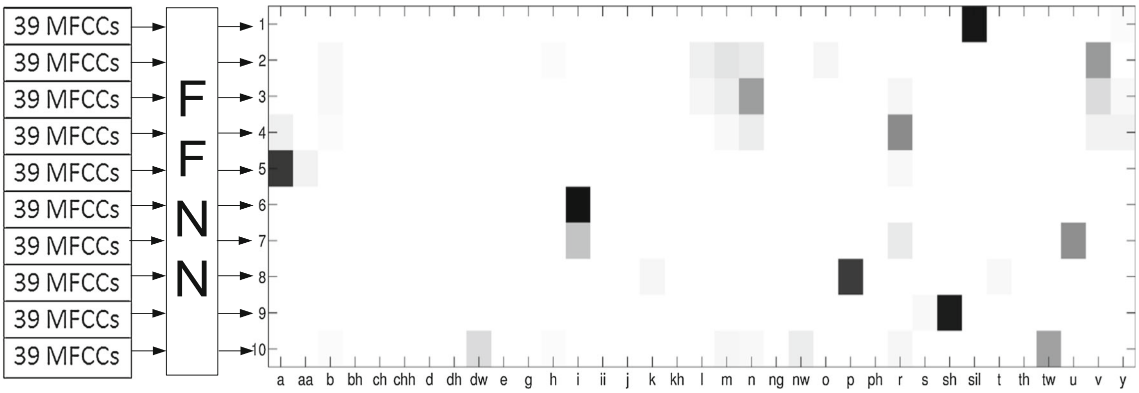

Fig. 3 Illustration of prediction of phone posteriors for ten frames using posteriogram representation

5.1 Development of baseline and 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 are 35, 31 and 31, respectively. Most frequently occurring phones in the IPA transcription are considered for building PRSs. HMM-based PRSs are developed using the procedure mentioned in Sect. 4.1.3. The baseline PRSs are developed using MFCCs as features. The detailed description of development of PRSs is given in Manjunath and Sreenivasa Rao (2014). The most common approach used to improve recognition accuracy of PRSs is to develop tandem systems (Hermansky et al. 2000). The AFs for each AF group are predicted from the spectral features using the FFNNs, as per the procedure mentioned in Sect. 4.1.4. 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 (Hermansky et al. 2000). 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 4 shows the block diagram of manner AF-based tandem PRS. Manner AFs are predicted using manner AF-predictor as shown in Fig. 1. 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 (Manjunath and Sreenivasa Rao 2015b; Manjunath et al. 2015a). 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. 4.1.5. Table 6 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 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.

5.2 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 (Sreenivasa Rao and Koolagudi 2013). The combined posterior

Fig. 4 Block diagram of the manner AF-based tandem PRS

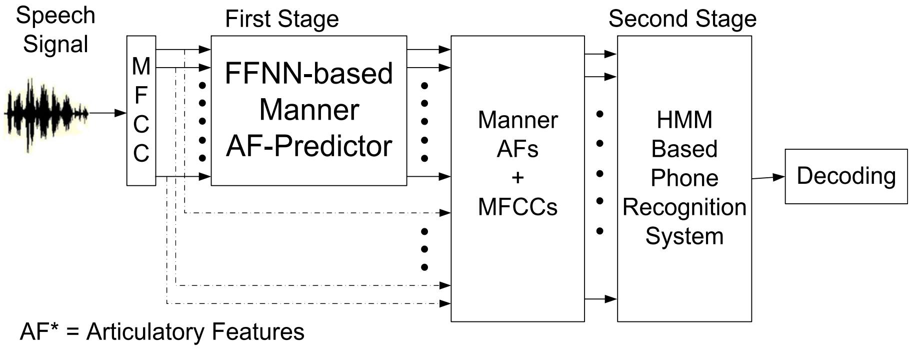
Speech Signal
First Stage
Second Stage
MFCC
FFNN-based Manner AF-Predictor
Manner AFs + MFCCs
HMM Based Phone Recognition System
Decoding
AF* = Articulatory Features

Table 6 Phone recognition accuracy (%) of AF-based tandem PRSs across read, extempore and conversation modes of speech

PRSs using different features	Recognition accuracy (%)		
	Read	Extemp	Conver
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

Extemp extempore, *Conver* conversation

probability P(j) of each frame with N phone classes, in the test utterance is given by the Eq. 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. $\sum_{i=1}^{k} w_i = 1$).

For each frame,

$$P(j) = \sum_{i=1}^{k} w_i * p_i(j), \quad \mathbf{j}\, \textit{varies from}\ 1\ \textit{to}\ N. \tag{4}$$

N—total number of phone classes; j—indicates specific phone class; k—number of PRSs considered for combining; i—indicates specific PRS.

Hybrid systems are developed by using the following combinations of AF-based tandem PRSs: (1) place and manner (2) roundness, frontness and height (3) 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. 4.2. The combination of MFCCs and PPs is used for developing PP-based tandem PRSs using HMMs.

Table 7 shows the optimal weighting factors used for developing hybrid PRSs of read, extempore and conversation modes of speech. 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 weighting factors w1, w2, w3, w4 and w5 correspond to place, manner, roundness, frontness and height AF-based tandem PRSs, respectively. We have explored all possible combinations of weighting factors, and the weights corresponding to optimal performance are treated as optimal weighting factors. The *hyphen* (-) symbol in Table 7 indicates that the particular weighting factor is not applicable for the corresponding hybrid PRS.

Further, we have developed PP-and-All-AF-based hybrid PRSs by combining PP-based tandem PRSs and All-AF-based hybrid PRSs in all three modes of speech. The weighting factor w1 corresponds to PP-based tandem PRSs while the weighting factor w2 corresponds to All-AF-based hybrid PRSs. 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. 4.1.5. Table 8 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

Table 7 The optimal weighting factors used for developing hybrid PRSs using weighted combination approach

Hybrid PRS	Weighting factors														
	Read					Extempore					Conversation				
	w1	w2	w3	w4	w5	w1	w2	w3	w4	w5	w1	w2	w3	w4	w5
Consonant-AF-based	0.5	0.5	–	–	–	0.6	0.4	–	–	–	0.5	0.5	–	–	–
Vowel-AF-based	–	–	0.3	0.3	0.4	–	–	0.1	0.4	0.5	–	–	0.4	0.2	0.4
All-AF-based	0.3	0.2	0.2	0.1	0.2	0.3	0.2	0.1	0.1	0.3	0.4	0.1	0.1	0.1	0.3
PP-and-All-AF-based	0.3	0.7	–	–	–	0.2	0.8	–	–	–	0.4	0.6	–	–	–

Table 8 Phone recognition accuracy (%) of hybrid PRSs across read, extempore and conversation modes of speech

PRSs using different features	Recognition accuracy (%)		
	Read	Extemp	Conver
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

Extemp extempore, *conver* conversation

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 improved recognition accuracy 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.

6 Phone recognition systems using excitation source and vocal tract system features

The excitation source information from LP residual is parameterized into two feature sets, namely: (1) residual Mel frequency cepstral coefficients (RMFCCs) and (2) Mel power differences of spectrum in sub-bands (MPDSS). RMFCCs and MPDSS features are extracted as per the procedure mentioned in Manjunath and Sreenivasa Rao (2015a). PRSs are developed using HMMs as described in 4.1.3. The phone recognition accuracy is determined using optimal string matching algorithm. Table 9 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 9, it is observed that the use of excitation source information resulted in 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 MPDSS features alone. The combination of MFCCs, RMFCCs and MPDSS have shown highest recognition accuracy in all three modes of speech. This is shows that the combination of vocal tract and excitation source information helps 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

Table 9 Phone recognition accuracy (%) of PRSs developed using excitation source and vocal tract system features across read, extempore and conversation modes of speech

PRSs using different features	Recognition accuracy (%)		
	Read	Extemp	Conver
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

extemp extempore, *conver* conversation

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.

7 Analysis of phone recognition across read, extempore and conversation modes of speech

The analysis of phone recognition across read, extempore and conversation modes of speech is carried out at two levels namely, (1) At broad phonetic subgroup level and (2) At phone level. The following subsections provide the detailed description of analysis of phone recognition across three modes of speech.

7.1 Analysis at broad subgroup level

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 (semi-vowels). The performance of PRSs using five subgroups is shown in Table 10. 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 10, 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 combination of MFCCs and excitation source features. The highest improvement obtained at subgroup level for read, extempore and conversation modes of speech is 3.35, 4.47 and 2.49 %, respectively. Although the PRSs based on five broad phonetic subgroups can't be directly used for developing speech recognition systems, but they can be used as a first-level classifiers in some applications like 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 recognition accuracy of fricatives and nasals is better with

Table 10 Phone recognition accuracy (%) of PRSs by considering five broad phonetic subgroups

PRSs using different features	Recognition accuracy (%)		
	Read	Extemp	Conver
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

extemp extempore, *conver* conversation

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, plosives. This kind of combination will lead to much better improvement at subgroup level. Semi-vowels have lowest classification accuracy. The misclassification mainly exists between vowels and semi-vowels. However, the misclassification of vowels into semi-vowels 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 the systems 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.

Further, the performance of PRSs is analyzed in all three modes of speech by merging all the unaspirated consonants to aspirated consonants. Table 11 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 combination of MFCCs and excitation source features.

From Table 11, it can be found that the improvement in the performance, after merging the unaspirated consonants to aspirated consonants, is higher in baseline PRSs compared to the PRSs using 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 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 differ 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.

Table 11 Improvement in the performance of PRSs across read, extempore and conversation modes of speech after merging unaspirated and aspirated consonants

PRSs using different features	Recognition accuracy (%)		
	Read	Extem	Conver
MFCCs (baseline)	1.62	1.59	2.24
MFCC + RMFCC + MPDSS	1.22	0.49	1.63

extemp extempore, *conver* conversation

7.2 Analysis at phone level

In this section the analysis of phone recognition is carried out at phone level. 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 majority of errors in extempore speech are 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: (1) Speakers have a tendency to use certain words of other language such as English, while having a conversation in Bengali. (2) Speakers speak very fast in conversation such that all the perceived sound units can not be located in the speech signal. (3) 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 semi-vowel 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 recognised. 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 recognised. For example $\{\langle silence\rangle\ kibhabe\}$ will be recognized as $\{\langle silence\rangle\ ibhabe\}$. 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 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 recognised, which is a major source of errors in conversation mode 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 better than conversation speech, the recognition accuracy of extempore speech is better than that of conversation speech.

8 Summary and conclusions

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 lead 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.

Acknowledgments The work presented in this paper was performed at IIT-Kharagpur as a part of the project 11(6)/2011-HCC(TDIL) , Dt. 23-12-2011, "Prosodically guided phonetic engine for searching speech databases in Indian languages" supported by Department of Information Technology, Government of India.

References

Bourlard, H. A., & Morgan, N. (1994). *Connnectionist speech recognition: A hybrid approach*. Dordrecht: Kluwer.

Chengalvarayan, R. (1998). On the use of normalized LPC error towards better large vocabulary speech recognition systems. In *IEEE international conference on acoustics, speech and signal processing* (pp. 17–20).

Dhananjaya, N., Yegnanarayana, B., & Suryakanth, V. G. (2011). 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).

Fallside, F., Lucke, H., Marsland, T. P., O'Shea, P. J., Owen, M. S. J., Prager, R. W., et al. (1990). Continuous speech recognition for the TIMIT database using neural networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 445–448).

Gerfen. (2015). *Phonetics theory* (online). http://www.unc.edu/gerfen/Ling 30Sp2002/phonetics.html.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5).

He, J., Liu, L., & Palm, G. (1996). On the use of residual cepstrum in speech recognition. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 5–8).

Hermansky, H., Ellis, D. P. W., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1635–1638).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*, 82–97.

Ketabdar, H., & Bourlard, H. (2008). 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).

Kirchhoff, K., Fink, Gernot A., & Sagerer, Gerhard. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, *37*, 303–319.

Lee, K., & Hon, H. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *37*, 1641–1648.

Manjunath, K. E., & Sreenivasa Rao, K. (2014). Automatic phonetic transcription for read, extempore and conversation speech for an indian language: Bengali. In *IEEE national conference on communications (NCC)* (pp. 1–6).

Manjunath, K. E., & Sreenivasa Rao, K. (2015a). Source and system features for phone recognition. *International Journal of Speech Technology*, *18*, 257–270.

Manjunath, K. E., & Sreenivasa Rao, K. (2015b). Improvement of phone recognition accuracy using articulatory features. *Applied Soft Computing* (revision submitted).

Manjunath, K. E., Sreenivasa Rao, K., & Gurunath Reddy, M. (2015a). Two-stage phone recognition system using articulatory and spectral features. In *IEEE international conference on signal processing and communication engineering systems (SPACES)* (pp. 107–111).

Manjunath, K. E., Sreenivasa Rao, K., & Gurunath Reddy, M. (2015b). Improvement of phone recognition accuracy using source and system features. In *IEEE international conference on signal processing and communication engineering systems (SPACES)* (pp. 501–505).

Manjunath, K. E., Sreenivasa Rao, K., & Pati, D. (2013). Development of phonetic engine for Indian languages: Bengali and Oriya. In *16th International oriental COCOSDA conference (IEEE explore)* (pp. 1–6), Gurgoan, India.

Manjunath, K. E., Sunil Kumar, S. B., Pati, D., Satapathy, B., & Sreenivasa Rao, K. (2013). Development of consonant-vowel recognition systems for Indian languages: Bengali and Oriya. In *IEEE INDICON (IEEE Explore)* (pp. 1–6), IIT Bombay, Mumbai, India.

Metze, F. (2005). *Articulatory features for conversational speech recognition*. Ph.D. dissertation, Carnegie Mellon University.

Mitra, V., Wang, W., Stolcke, A., Nam, H., Richey, C., Yuan, J., et al. (2013). Articulatory trajectories for large-vocabulary speech recognition. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 7145–7149).

Mohamed, A., Dahl, G. E., & Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*, 14–22.

Sainath, T. N., Mohamed, A., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. In *IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 8614–8618).

Siniscalchi, S. M., & Lee, C. (2009). A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. *Speech Communication*, *51*, 1139–1153.

Speech Group at the International Computer Science Ins. (2010). *QuickNet software and documentation* (online). http://www1.icsi.berkeley.edu/Speech.

Sreenivasa Rao, K., & Koolagudi, S. G. (2013). Recognition of emotions from video using acoustic and facial features. In *Signal, image and video processing (SIViP)* (pp. 1–17).

Sunil Kumar, S. B., Sreenivasa Rao, K., & Pati, D. (2013). Phonetic and prosodically rich transcribed speech corpus in indian languages: Bengali and Odia. In *16th International Oriental COCOSDA* (pp. 1–5).

The Hidden Markov Model Toolkit and HTK book. (2015). (online). http://htk.eng.cam.ac.uk.

The International Phonetic Association. (2015). International Phonetic Alphabet (online). http://www.langsci.ucl.ac.uk/ipa/index.html.

Toth, L. (2014). Convolutional deep maxout networks for phone recognition. In *International speech communication association (INTERSPEECH)* (pp. 1078–1082).