Circuits Syst Signal Process (2018) 37:704–728
https://doi.org/10.1007/s00034-017-0568-8

Improvement of Phone Recognition Accuracy Using Articulatory Features

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

Received: 4 February 2016 / Revised: 21 April 2017 / Accepted: 25 April 2017 /
Published online: 8 May 2017

Abstract In this work, we have explored the articulatory features (AFs) for improving the performance of the phone recognition systems (PRSs) using TIMIT and Bengali speech corpora. AFs for manner, place, roundness, frontness, and height groups are derived from the spectral features using feedforward neural networks. MFCCs are used as spectral features. HMMs and DNNs are explored for developing the PRSs. The combination of MFCCs and AFs is used to develop tandem PRSs. Five tandem systems based on manner, place, roundness, frontness, and height AFs are developed. We have also developed a phone-posterior-based tandem system using the phone posteriors derived from the MFCCs through feedforward neural networks. The tandem systems are then combined to develop hybrid systems using weighted combination scheme. A systematic analysis of phone-level accuracies contributed by individual AF groups, consonant AF groups, and vowel AF groups is carried out separately. The combination of all the AF-based tandem PRSs and phone-posterior-based tandem PRS has shown highest phone recognition accuracy for both Bengali and TIMIT datasets. DNNs have outperformed HMMs in all the cases. The best performing systems have shown recognition accuracy of 55.8% and 74.7% for Bengali and TIMIT datasets, respectively.

Keywords Phone recognition · Articulatory features · Phone posteriors · Bengali · Hybrid PRS · Tandem PRS · FFNNs · HMMs · DNNs

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

K. Sreenivasa Rao
ksrao@iitkgp.ac.in

[1] Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

1 Introduction

Phone recognition generally involves building phone models using pattern recognition techniques such as Hidden Markov Models (HMMs), FeedForward Neural Networks (FFNNs), Support Vector Machines (SVMs) and Deep Neural Networks (DNNs). Generally, the standard spectral features such as Linear Prediction Cepstral Coefficients (LPCCs) or Mel-Frequency Cepstral Coefficients (MFCCs) are used for developing phone recognition models. Spectral features mainly represent the gross shape of the vocal tract system in terms of dominant resonant frequencies during production of a sound unit. But, the production of a sound unit depends on both the gross shape of the vocal tract as well as the positioning and movements of various articulators. The positioning and movements of articulators during the production of a sound unit can be represented using articulatory features (AFs). The performance of the phone recognition systems (PRSs) can be significantly improved with the use of AFs along with the spectral features [5,11,22,26]. Motivated by this, in this study, we have explored the combination of the spectral features and the AFs with an intent to improve the performance of PRSs.

Speech is produced by the exhalation of air from the lungs leading to the vibration of the vocal folds followed by passing of air through the vocal tract and then radiating out through the nostrils or lips. The articulators such as lips, teeth, tongue, alveolar ridge, hard palate, velum and glottis are involved in the speech production. 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 [46]. 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. The continuous-valued measurements are used in [10,49,52], while the discrete-valued measurements are used in [5,11,22]. In this work, the AFs are represented using discrete values. For example, the discrete values for roundness AF group are rounded and unrounded [21]. The significance of having five AF groups to capture various AFs is as follows: *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. 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. *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 manners of articulation. *Roundedness* indicates whether the lips are rounded or not during the production of a vowel. *Frontness* indicates the horizontal position of the tongue during the articulation of a vowel relative to the front

of the mouth. *Height* denotes the vertical position of the tongue during the production of a vowel relative to the aperture of the jaw [13].

Generally, the PRSs are developed using HMMs, FFNNs, SVMs and DNNs. HMMs are the generative models, and they can be used to model the sequence of vocal tract shapes contributed to the production of a sound unit, with local spectral variability modeled using the mixtures of Gaussian densities [38]. The FFNNs are the discriminative classifiers, and they have good discriminative power to distinguish between the correct output class and the rival ones [25]. In the context of phone recognition, discrimination between the vocal tract shapes offered by various sound units is exploited by the FFNNs. DNNs have many hidden layers, which can model more complex nonlinear relationships [17]. The use of DNNs has lead to dramatic improvement in the performance of speech recognition systems in the recent years. In this work, we have used HMMs and DNNs for developing PRSs and the FFNNs for deriving tandem features. The main focus of this study is to improve the performance of the PRSs using the AFs. The AFs act as additional clues, which aid in discriminating between various sound units. The AFs provide supplementary information, which can be used along with the spectral features to improve the performance of the PRSs.

The rest of the work is organized as follows: Sect. 2 provides the literature survey. Section 3 describes the speech corpora used in this study. Section 4 discusses different types of feature extraction techniques used in this work. Section 5 describes the development of the baseline and tandem PRSs. Section 6 provides details of the development of hybrid PRSs using the weighted combination scheme. Section 7 discusses about the comparison of the proposed PRSs with other related works. Section 8 provides the summary and conclusion of the paper.

2 Literature Survey

The area of speech recognition has been one of the most active areas of research for the last six decades. The most common approaches of developing speech recognition systems use HMMs [23], FFNNs [9], hybrid systems using combination of HMMs and FFNNs [4], and DNNs [15,17,31]. From the existing literature, it is observed that there are some studies exploring the AFs to improve the phone recognition accuracy.

Few works related to development of speech recognition systems using discrete-valued AFs are as follows: In 2009, Siniscalchi et al. have developed automatic speech recognition systems using acoustic-phonetic information. The acoustic-phonetic information is derived using the lattice rescoring approach. A bank of speech event detectors are used to score the place and manner of articulation events. Three speech recognition tasks namely continuous phone recognition, connected digit recognition and large vocabulary continuous speech recognition are carried out using the lattice scoring approach. The lattice rescoring framework has achieved better results in all three cases [43]. In 2007, Cetin et al. have used the AFs to develop a tandem speech recognition system. The AFs are derived by training multilayer perceptrons using the spectral features. Fisher and the SVitchboard speech corpora are used for evaluating the prediction accuracy of AFs. The derived AF evidences along with perceptual linear prediction features are used for improving the word error rate [5,11]. In 2002, Kirchhoff et al.

have used AFs to develop robust speech recognition systems. Two different recognition tasks namely continuous digit recognition in case of 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, and the combination of acoustic features and AFs consistently leads to a significant reduction in the word error rate across all acoustic conditions [22].

Some of the works exploring the continuous-valued AFs to improve the performance of speech recognisers are listed below. In 2007, Frankel et al. have used the linear dynamic models to improve the performance of HMM-based speech recognisers. The internal variables of the hidden states are captured through linear dynamic models. These internal variables reflect the properties of slowly and continuously moving articulators along highly constrained trajectories. It is shown that the use of linear dynamic models has resulted in significant improvement in the performance [10,12]. In 2011, Ghosh et al. have estimated the articulatory features using subject-independent acoustic-to-articulatory inversion. It is found that the inclusion of the articulatory information improves classification accuracy. The improvement is more significant if the speaking style of the exemplar matches with that of the talker. AFs are specified in terms of five vocal tract constriction variable trajectories called tract variables (TVs), namely lip aperture, lip protrusion, jaw opening, tongue tip constriction degree, and tongue body constriction degree [14]. In 2013, Mitra et al. have estimated the TVs from the speech signal using artificial neural networks. Eight TVs related to constriction of lip, tongue tip, tongue body, velum, and glottis are captured. It is found that the combination of MFCCs and TVs yields higher recognition accuracy [26]. TVs are explored for robust speech recognition in [27] and [28].

In most of the above works, the AFs are mostly used as tandem features to improve the recognition accuracy of PRSs. Hence, we have explored the weighted combination scheme for combining the AFs from various AF groups. The hybrid PRSs are developed using the weighted combination of various AFs. In this work, we have carried out a systematic analysis of the phone-level accuracies contributed by each AF group. The analysis is carried out by developing separate hybrid PRSs based on consonant AFs and vowel AFs. The consonant-based hybrid PRSs are developed using the *place* and *manner* AFs, whereas the vowel-based hybrid PRSs are developed using the *roundness*, *frontness* and *height* AFs. The existing studies have mostly explored either the combination of spectral features and the phone posteriors, or the combination of spectral features and the AFs, separately to improve the performance of the PRSs. But, in this work, we have proposed the combination of these two systems, to further enhance the performance of the PRSs. 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, we have explored the AFs in the context of an Indian language Bengali. The use of DNNs to develop AF-based PRSs is not much explored in the literature. Hence, we have explored DNNs in addition to HMMs to develop AF-based PRSs. Since the AFs provide supplementary information for phone recognition, the combination of spectral and the articulatory features leads to significant improvement in the performance of the PRSs and the same is widely reported in the literature [5,11,26–29]. The objective of our study is to examine the role of AFs in improving the performance of PRSs.

3 Speech Corpora

For developing and analyzing the performance of the 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 Bengali speech corpus and TIMIT English speech corpus are discussed in the following subsections.

3.1 Bengali Speech Corpus

The PPRT Bengali speech corpus developed at IIT Kharagpur is used in this study [44]. The speech corpus contains the speech data collected in three different modes namely read mode, extempore mode and conversational mode. Speech wave files are sampled at 16 kHz, and each sample is encoded using 16 bits. The speech data in all three modes of speech are transcribed using the IPA chart. The IPA chart 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 far fewer symbols [46]. In our case, we were able to represent the speech utterances in Bengali language with 64 IPA symbols plus one *hyphen* used for indicating silence. We have organized the speech data in the form of sentences to conduct experiments. The duration of the speech data used in this study is about 1.16 h of read speech spoken by 13 female speakers and 8 male speakers. The data used for training and testing were from different speakers. For training, around 80% of the data was used and remaining 20% of the data was used for testing. 10% of training data is used as development set. Each speaker in the training set had equal contribution toward development set. The development set is used to determine the learning rates and to decide on when to terminate training. We have considered overlapped development set for HMMs, and non-overlapping held-out development set for FFNNs and DNNs. In case of overlapped development set, the development set is used for both training and cross-validation.

3.2 TIMIT Speech Corpus

TIMIT speech corpus is a read speech corpus designed for its use in acoustic-phonetic studies. TIMIT is widely used in the development and evaluation of automatic speech recognition systems. TIMIT corpus contains 16 bit, 16 kHz speech wave files along with the time-aligned orthographic, phonetic and word transcriptions for each utterance. The corpus was jointly designed by Massachusetts Institute of Technology, SRI International and Texas Instruments. The transcriptions in TIMIT are hand verified [18]. The training set and core test set, as suggested in the 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 3–5 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.

4 Feature Extraction

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

4.1 Mel-Frequency Cepstral Coefficients

The MFCC features mainly capture the vocal tract information. The following procedure is used for extracting the MFCCs from the speech signal. The speech signal is divided into frames with a duration of 25 ms [38]. A frame shift of 10 ms is employed for locating the adjacent frames. The frames are Hamming windowed to reduce the edge effect, while taking the Discrete Fourier Transform on the signal. For each frame, cepstral coefficients are computed using a Mel filter bank with 26 Mel filters. The speech is parameterized into 13 MFCCs including 0^{th} cepstral coefficient as well as their first- and second-order derivatives resulting to a total of 39 components.

4.2 Extraction of Articulatory Features

In this study, we have considered five AF groups, namely: place, manner, frontness, roundness and height. The prediction of the AFs from the spectral features using FFNNs is discussed in detail in the following subsections.

4.2.1 Articulatory Features

The AFs provide the 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. The 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 finer variations in the shape of the vocal tract. The co-articulation effect between the adjacent phonetic units is captured by the AFs [8,26,34,37]. The AFs provide additional clues for discriminating among the various sound units.

There are mainly three ways to derive AFs: (i) acoustic-articulatory transformations using inverse mapping, (ii) direct physical measurements, and (iii) classification scores for pseudo-articulatory features [22]. In the first approach, the articulatory movements are estimated from the speech acoustics through inverse mapping of speech. The process of inverse mapping refers to the inverse of the natural transformation from articulatory movements to speech acoustics [7,47]. In [6], the frication and voicing features are detected using zero-frequency filtered signal. Since there are inverse fil-

Table 1 Articulatory feature specification for Bengali and TIMIT datasets

AF group (Cardinality)	Features
	Bengali
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
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

tering methods for detecting only some of the AFs and not for all the AFs, the use of inverse mapping technique will be difficult in case of continuous speech recognition applications.

Second approach deals with capturing the motions of articulators through direct physical measurement techniques such as X-ray filming (cineradiography), magnetic resonance imaging [32,33], electromagnetic articulography [24,45], and electropalatography [3]. Use of these techniques requires costly setup and involves the risk of health hazards such as exposure to radiations. Moreover, the speech corpora with physical measurements of articulatory motions are not available for Indian languages. The works based on physical measurement techniques are reported in [1,14] and [29]. In the third approach, AFs are derived from the acoustic signal using statistical classifiers. The spectral features from the acoustic signal are given as input, and the classification scores are obtained at the output of the classifier. The classification scores thus obtained indicate pseudo-articulatory features. Among three approaches, third approach is more feasible and popularly used [5,11,22,39]. Hence, in this work, we have explored the third approach.

Although most of the existing works use spectral features such as MFCCs or PLPs to derive the AFs, but there are few works exploring *Mel-log filter bank* features to estimate AFs. In [30], Mizera et al. have shown that the use of *Mel-log filter bank* features for estimating the AFs has better results than MFCCs or PLPs. They have computed the *Mel-log filter bank* features using 23 bands of log mel-scaled filter bank in the range 64–8000 Hz.

We have captured the discrete information about the positioning and movement of articulators with respect to five AF groups. Each AF group along with their possible AF values is shown in Table 1. Table 1 shows the specification of AFs 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 of *manner* AF group are same for both Bengali and TIMIT datasets, while the possible feature values for remaining AF groups are different for Bengali and TIMIT datasets. The difference in the feature specification of *place* AF group is due to *retroflex* feature value. In IPA chart, the phoneme /r/ (based on its production characteristics) is represented by five different symbols {/r/, /ɾ/, /ɹ/, /ɽ/, /ɻ/} [46]. Among five symbols, {/r/, /ɾ/, /ɹ/} are approximants, and {/ɽ/, /ɻ/} are retroflexes. In TIMIT transcription /r/ is represented using {/er/, /r/} symbols. There is no way to determine whether a /r/ is an approximant or a retroflex using TIMIT transcription. Hence, all the phones {/er/, /r/} of TIMIT are mapped to approximant but not retroflex. Hence, there will be no separate class for retroflexes in TIMIT. We have derived this mapping based in the previous works reported in [6] and [39], which use TIMIT dataset. The studies reported in [5] and [11] also map the phones {/er/, /r/} to approximant but not to retroflex. SVitchboard and telephone speech corpora are used in [11] and [5], respectively. In case of Bengali, the transcription is derived using IPA chart. Hence, we have the flexibility to determine whether a phone /r/ is an approximant or a retroflex using IPA transcription. In addition to retroflex /r/, other retroflex symbols such as /T/ and /D/ are found in abundance in Bengali transcription. Hence, it is possible to have a separate class for retroflex in Bengali. This results in having a retroflex feature value for *place* AF group of Bengali.

The difference in the feature specification of *roundness*, *frontness* and *height* AF groups is due to *diphthong* feature value. In IPA transcription, diphthongs are transcribed as their component vowels (i.e., starting and ending vowel) [2]. Since Bengali uses the IPA transcription, there was no separate representation indicating diphthongs explicitly in Bengali transcription. Hence, there was no separate representation (i.e., separate feature value) for indicating diphthongs explicitly in Bengali transcription. In case of TIMIT dataset, diphthongs are transcribed using separate phonetic symbols. Since the production characteristics (*i.e., frontness, roundness, and height*) of two different vowels of a diphthong are not same, we cannot represent the diphthong using the feature values considered for *roundness*, *frontness* and *height* AFs.

4.2.2 Prediction of Articulatory Features

In this work, the 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 the AF-predictors. Figure 1 shows the block diagram for the prediction of manner AFs. HMM and FFNN-based AF-predictors [5,11] are developed for the manner AF group using MFCCs. The predicted feature values represent the manner AFs.

Figure 2 illustrates the prediction of manner AFs for ten frames using the 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 bands in the posteriogram indicate higher posterior probability, while the lighter bands indicate lower posterior probability. The labels on 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 distribu-

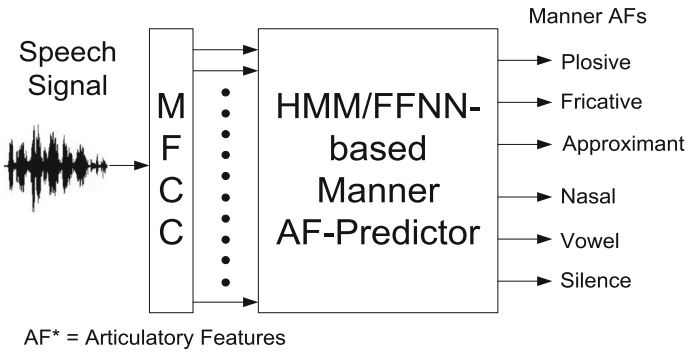

Fig. 1 Block diagram of manner articulatory features predictor

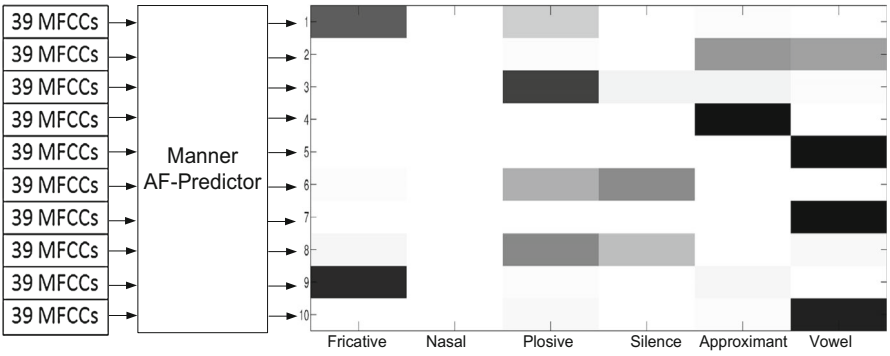

Fig. 2 Illustration of prediction of manner articulatory features for ten frames using posteriogram representation

tion 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.

Similarly, the AF-predictors are developed for the remaining four AF groups, as shown in Fig. 3. The AFs for a particular AF group are predicted using the AF-predictor of that specific group.

4.2.2.1 Mapping Phone Labels to AF Labels The AF-predictors are developed by training HMMs and FFNNs. For training HMMs and FFNNs, we require the speech data which is transcribed at AF-level. The AF-level transcription indicates the transcription derived using the AF labels. Since the transcription is available at phone level, we derive the transcription at AF-level by mapping the phone labels in the phone-level transcription to AF labels. The 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 1. 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 2 and 3, respectively. First column in Table 2 lists the unique IPA symbols used in Bengali transcription, while the first column in Table 3 lists the unique phones used in TIMIT transcription. Second to sixth columns show the corresponding place, manner, roundness, frontness and height AF

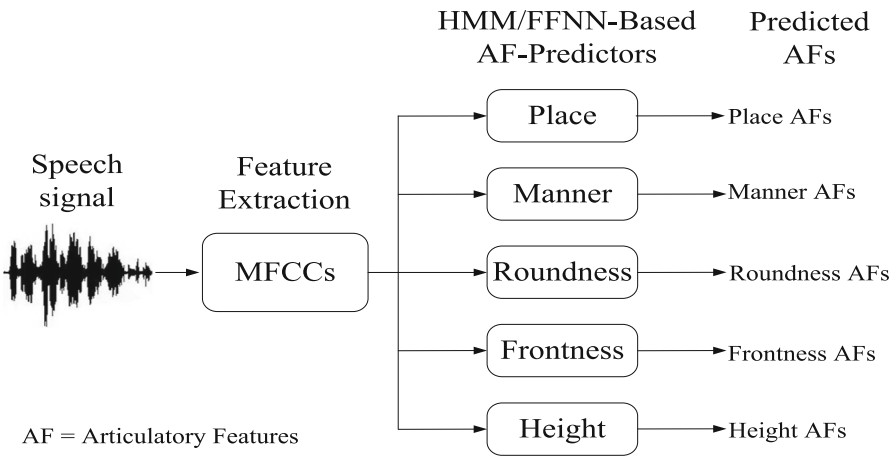

Fig. 3 Block diagram for the prediction of articulatory features

values, respectively, for each phone. The mapping for Bengali dataset is derived using the IPA chart [46], whereas the mapping for the TIMIT dataset is derived with the aid of *TIMIT to IPA mapping* as shown in [42].

4.2.2.2 Development of AF-Predictors Using HMMs 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 sound unit. The embedded reestimation using Baum–Welch training is followed by the Viterbi decoding of the test utterances. The development set is used to tune the system parameters such as number of iterations, word-insertion penalty, and grammar scale factor. The open-source HTK toolkit is used for building HMM models [50].

4.2.2.3 Development of AF-Predictors Using FFNNs 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 each sound unit, the MFCC feature vectors are fed to input layer and the information about the AF label is given at the output layer during training of the neural network. Three-layered FFNN with sigmoid nonlinearity at the hidden layer, and softmax nonlinearity at the output layer is used. During training, multiple passes are made through the entire set of the 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 the development set. The training process will be stopped after the epoch at which the increment in the performance improvement is less than 0.5% with development set. The advantage of *development set*-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 proba-

Table 2 Mapping of phone labels to AF groups in Bengali 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
r ɹ ɾ l	alveolar	approximant	nil	nil	nil
ɭ	retroflex	approximant	nil	nil	nil
ʋ	labiodental	approximant	nil	nil	nil
sil	silence	silence	silence	silence	silence

bilities 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 the posterior probabilities of each class [35]. The posterior probabilities are used for representing the AFs of a sound unit. The open-source quicknet software is used for training the FFNNs [48].

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. Most of the existing works have used a

Table 3 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

temporal context of nine frames [5,11,19]. But, when we experimented with various context sizes, it is found that the context of three frames performs better compared to all other contexts. This might be due to the lower amount of training data used in this study compared to the amount of training data used in [5,11,19]. It is found that [5,11] use close to 2000 h of training data, and [19] uses 16 h of training data. Hence, we have used a context of three frames in this study. The temporal context of 3 frames 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 by using Eq. 1.

Table 4 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

$$NNIL = \text{No. of frames in temporal context} \times \text{No. of MFCCs per frame} \quad (1)$$

According to Eq. 1, the number of nodes in input layer will be 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 the 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 1. Table 4 shows the number of epochs carried out during training of FFNNs for various AF groups of the 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.

4.2.3 Performance Evaluation of AF-Predictors

The accuracy of the AF-predictors is determined by comparing the decoded AF labels with the reference AF labels by performing an optimal string matching using dynamic programming [50]. Once the optimal alignment is found, the number of substitution errors (S), deletion errors (D) and insertion errors (I) is determined. The recognition accuracy in percentage is calculated using Eq. 2. The terms *recognition accuracy* and *phone recognition accuracy* are used interchangeably. The recognition accuracy of the AF-predictors is termed as *prediction accuracy*.

$$Recognition\ Accuracy = \frac{N\text{-}D\text{-}S\text{-}I}{N} \times 100\% \quad (2)$$

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

Classification accuracy of a class label is defined as the ratio of number of correctly classified samples of the class label to the total number of samples of that class label. Equation 3 gives the formula for computing the classification accuracy.

$$Classification\ Accuracy = \frac{\text{Number of samples correctly classified}}{\text{Total number of sample cases}} \times 100\% \quad (3)$$

Table 5 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

Table 5 Prediction accuracy of AF-predictors of different AF groups

AF group	Prediction accuracy (%)			
	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

show the prediction accuracies for Bengali dataset, while the fourth and fifth columns tabulate the prediction accuracies for TIMIT dataset. The results are shown separately for HMM-based and FFNN-based systems. It can be observed that the prediction accuracies of all the AF groups are higher in case of FFNNs compared to HMMs for Bengali dataset, while the prediction accuracies of most of the AF groups are higher in case of FFNNs compared to HMMs for the TIMIT dataset. Though 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 prediction accuracies for all AF groups of Bengali dataset and for majority of AF groups in TIMIT dataset, we have used the FFNNs for predicting the AFs of various AF groups. Since the FFNNs provide a discriminative way of estimating posterior probabilities [20], it is more advantageous to use FFNNs for developing AF-predictors than HMMs. The combination of the discriminative knowledge captured by the AF-predictors and the sequential knowledge captured by the HMMs (during the development of PRSs) will lead 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 different AF groups.

Place Labiodentals have poor classification accuracy. Labiodentals are misclassified into bilabials. Retroflexes are misclassified into alveolars. All the groups have significant misclassifications into alveolars. Alveolars and velars have more deletion errors.

Manner Plosives have very poor classification accuracy, due to their misclassifications into nasals. Plosives are also misclassified into silence, due to unvoiced plosives such as p,t,k classified as silence. Vowels have highest classification accuracy.

Roundness Unrounded to rounded misclassification is more prominent. Consonants are mainly misclassified into vowels and have more deletion errors.

Frontness Mid has least classification accuracy, due to its misclassification into back. Consonants are mainly misclassified into vowels and have more deletion errors.

Height Mid-high has least classification accuracy. High to mid-high and mid-low to mid-high misclassifications are prominent. Consonants are mainly misclassified into mid-low.

We have also computed the framewise accuracies of FFNN-based AF-predictors as shown in [5,11,39]. The framewise accuracies of FFNN-based AF-predictors are shown in Table 6. The results are computed for both training and development sets.

Table 6 The framewise accuracies of each FFNN-based AF-predictors on training and development datasets

AF group	Bengali		TIMIT	
	Train	Development	Train	Development
Place	83.97	83.33	83.55	82.77
Manner	87.91	87.00	87.45	86.72
Roundness	90.06	88.69	86.82	85.84
Frontness	88.90	88.21	85.35	84.03
Height	83.45	82.04	81.59	80.20

5 Development of Baseline and Tandem Phone Recognition Systems

The purpose of developing a PRS is to convert the speech signal into a sequence of basic sound units, namely phones. In this study, we have developed Bengali and English PRSs using HMMs and DNNs. Most frequently occurring phones in the IPA transcription are considered for building Bengali PRS. The number of phones considered for developing Bengali PRSs is 35. The 61 phones of TIMIT dataset are downsized to 48 phones by using the approach shown in [23]. TIMIT PRSs are trained using 48 phones and tested using 39 phones as described in [23]. The PRSs are developed without using any language related information, i.e., no language model is used. HMM-based PRSs are developed using the procedure mention in Sect. 4.2.2.2.

The procedure for developing PRSs using DNNs is as follows: DNNs with 3 hidden layers having tanh nonlinearity at hidden layers and softmax activation at the output layer are used. DNNs are trained using greedy layer-by-layer supervised training. Initial learning rate was chosen to be 0.015 and was decreased exponentially for the first 15 epochs. A constant learning rate of 0.002 was used for the last 5 epochs. Mixing up was carried out in the halfway between the completion of addition of all the hidden layers and the end of training. Preconditioned affine components are used to maintain the stability of the training. The final model is obtained by combining the models from last 10 iterations into a single model. Each input to DNNs uses a temporal context of 9 frames (4 frames on either side).

The size of the input layer depends on the dimension of features used for training the DNNs. The language model weighting factor and acoustic scaling factor used for decoding the lattice are optimally determined using the development set to maximize the recognition accuracy. DNNs training used in this study is similar to the one discussed in Sect. 2.2 of [51], except that our setup had a single CPU. DNNs are built using the open-source speech recognition toolkit Kaldi [36].

The baseline PRSs are developed using MFCCs as features. We have developed AF-based tandem PRSs 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. 4.2.2.3. In tandem approach, the 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 features for developing PRSs. The predicted AFs of a particular AF group are augmented with the MFCCs to develop AF-based tandem PRS for that AF group [16]. Separate tandem PRSs are

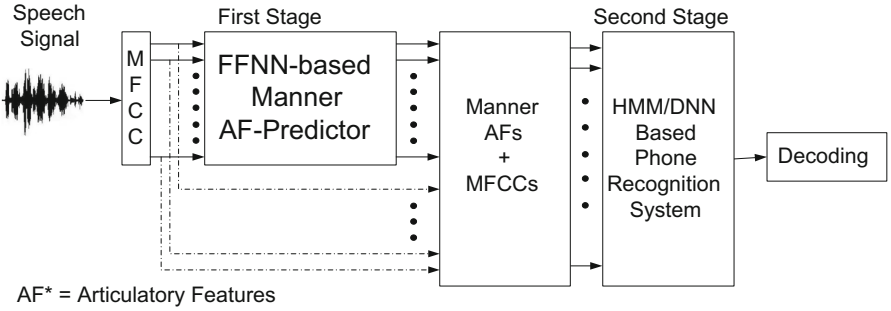

Fig. 4 Block diagram of the manner AF-based tandem PRS

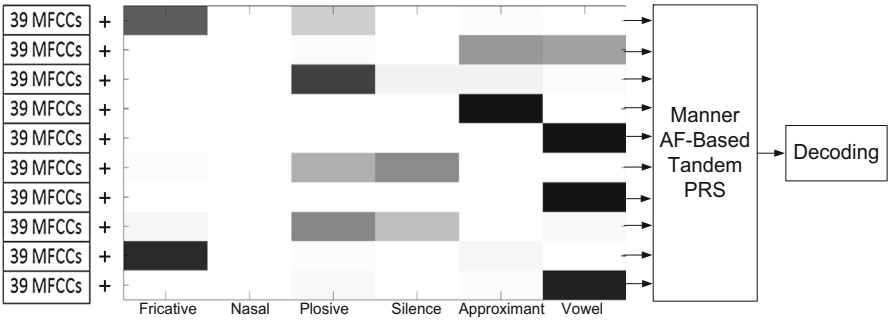

Fig. 5 Illustration of manner AF-based tandem PRS for ten frames using posteriogram representation

developed using the AFs predicted from each AF group. This leads to development of five different AF-based tandem PRSs. Figure 4 shows the block diagram of manner AF-based tandem PRS. The manner AF-predictor is used for predicting the manner AFs, as shown in Fig. 1. The combination of predicted manner AFs and MFCCs is used as features to develop HMM/DNN-based tandem PRS in the second-stage. Similarly, five different tandem PRSs are developed using the predicted AFs from each AF group.

Figure 5 illustrates the manner AF-based tandem PRS for ten frames using posteriogram representation. The MFCCs are augmented with the posteriogram distribution of the manner AFs obtained in first stage (shown in Fig. 2). The combination of the MFCCs and the manner AFs is then fed to the manner AF-based tandem PRS for decoding the phones in the input speech utterance.

The phone recognition accuracy is determined as per the procedure mentioned in Sect. 4.2.3. Table 7 shows the phone recognition accuracies of the baseline and the tandem PRSs for Bengali and TIMIT datasets. First column shows different types of features used in the development of PRSs. Second and third columns indicate the recognition accuracies of Bengali PRSs, while the last two columns show the performance of TIMIT PRSs. It can be observed that all tandem PRSs have higher recognition accuracies compared to their respective baseline PRSs. The combination of MFCCs and the *Height AFs* has shown highest recognition accuracy for the Bengali dataset, while the combination of MFCCs and the *Manner* (or *Roundness*) AFs has shown the highest recognition accuracy for the TIMIT dataset. It is observed

Table 7 Phone recognition accuracy of baseline and AF-based tandem PRSs

Features	Recognition accuracy (%)			
	Bengali		TIMIT	
	HMMs	DNNs	HMMs	DNNs
MFCCs	45.48	50.20	58.45	69.10
MFCCs + place AFs	48.89	51.40	60.93	70.00
MFCCs + manner AFs	47.74	50.60	61.43	69.70
MFCCs + roundness AFs	47.28	51.30	60.75	70.00
MFCCs + frontness AFs	46.59	50.60	61.11	69.80
MFCCs + height AFs	48.60	51.80	61.58	69.60

that the classification accuracy of the aspirated plosives is decreased in all the five AF-based tandem PRSs, whereas the classification accuracy of most of the unaspirated plosives, fricatives and approximants is increased in all the AF-based tandem PRSs compared to baseline system. The classification accuracy of *silence* has improved in all the AF-based tanden PRSs compared to the baseline system.

The analysis of each AF-based tandem PRS is as follows:

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

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

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

Frontness AF-based tandem PRS The classification accuracy of all the vowels is improved. The back vowels have shown the highest improvement, and the mid vowels have shown least improvement in their classification accuracies.

Height AF-based tandem PRS The classification accuracy of all the vowels is improved. The mid-low subgroup has shown the least improvement, and the mid-high subgroup has shown the highest improvement in its classification accuracy.

6 Hybrid Phone Recognition Systems Using Articulatory Features

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

6.1 Development of Hybrid Phone Recognition Systems Using Articulatory Features

The hybrid PRSs are developed by combining the AF-based tandem PRSs using weighted combination scheme. The following combinations of AF-based tandem PRSs are used in development of hybrid PRSs : (i) place and manner (ii) roundness, frontness, and height (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. The hybrid PRSs developed using roundness, frontness and height AF-based tandem PRSs are called Vowel-AF-based hybrid PRSs, as the roundness, frontness and height AFs mainly capture the characteristics of the vowels. The hybrid PRSs developed using the combination of all the five AF-based tandem PRSs are called All-AF-based hybrid PRSs.

Figure 6 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 Vowel-AF-based, Consonant-AF-based and All-AF-based hybrid PRSs are developed by combining the scores from the tandem PRSs using weighted combination approach. The details of the weighted combination scheme are discussed in Sect. 6.2.

The phone-posterior-based tandem PRSs are developed to compare the performance of the AF-based hybrid PRSs with the phone-posterior-based tandem PRSs. To develop phone-posterior based tandem PRSs, the FFNNs are first trained to perform the classification at frame level, and then the frame-level phone posterior estimates of the FFNNs are used as features for developing PRSs. The FFNNs are trained as per the procedure mentioned in Sect. 4.2.2. The MFCCs are fed at the input layer, and information about the phone label is fed at the output layer. 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

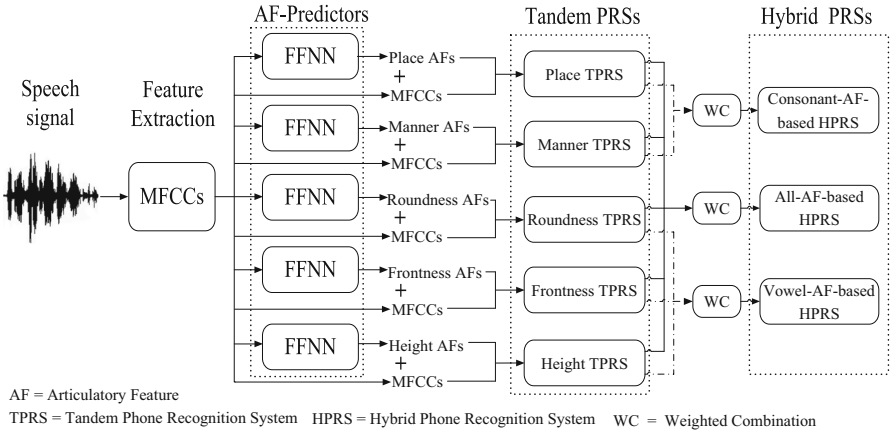

Fig. 6 Block diagram of hybrid phone recognition systems

considered for training FFNNs. The number of epochs carried out for TIMIT and Bengali is 8 and 13, respectively. The MFCCs are augmented with the phone posteriors derived from the FFNNs to develop phone-posterior-based tandem PRSs using HMMs or DNNs.

6.2 Fusion of Posterior Probabilities Using Weighted Combination Scheme

The posterior probabilities from different AF groups are combined using weighted combination scheme [41]. In weighted combination scheme, the posterior probabilities from different PRSs are combined at frame level. We have explored sum rule, product rule, min rule, and max rule for fusion of posterior probabilities from multiple streams [22]. It is found that the performance using sum rule is better than all others. Hence, we have considered sum rule for fusion of posterior probabilities in our study. The combined posterior probability $P(j)$ of each frame with N phone classes, in the test utterance is given by Eq. 4.

$$\begin{aligned} &\text{For each frame,} \quad P(j) = \sum_{i=1}^{k} w_i * p_i(j), \\ &\text{where,} \quad j \text{ varies from 1 } to\ N. \\ &N \ = \text{ Total number of phone classes.} \\ &j \ = \text{ indicates specific phone class.} \\ &k \ = \text{ Number of } PRSs \text{ considered } for \text{ combining.} \\ &i \ = \text{ indicates specific } PRS. \end{aligned} \tag{4}$$

The value of a weighting factor w_i (where i stands for ith PRS) is determined by varying from 0 to 1 with a step size of 0.1. This leads to eleven possible values for a weighting factor. They are as follows: *{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}*. Each PRS is associated with a weighting factor. If there are n PRSs to be combined then there will be n weighting factors. The collection of all the n weighting factors is called a *weighting factors set*. The phone recognition accuracy is determined for all the possible set of weighting factors using development set. The set of empirically determined best performing weights on the development set are considered as the *optimal weighting factors set*.

Table 8 shows the optimal weighting factors for different hybrid PRSs obtained using development set. First column lists different types of hybrid PRSs. Second to sixth columns indicate the weighting factors used for Bengali dataset, while the last five columns indicate the weighting factors used for TIMIT dataset. The *hyphen* (–) symbol in Table 8 indicates that the particular weighting factor is not applicable for 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 8 have shown the highest recognition accuracies on the development set. We have also combined the phone-posterior-based tandem PRS and the All-AF-based

Table 8 Optimal weighting factors used for developing hybrid PRSs determined using development set

HPRS	Bengali					TIMIT				
	w1	w2	w3	w4	w5	w1	w2	w3	w4	w5
Consonant-AF-based	0.6	0.4	–	–	–	0.5	0.5	–	–	–
Vowel-AF-based	–	–	0.4	0.4	0.2	–	–	0.4	0.2	0.4
All-AF-based	0.3	0.1	0.2	0.2	0.2	0.3	0.2	0.1	0.2	0.2
PP-and-all-AF-based	0.2	0.8	–	–	–	0.3	0.7	–	–	–

Table 9 Phone recognition accuracy (%) of phone-posterior-based and AF-based hybrid phone recognition systems using Bengali and TIMIT datasets

PRSs	Recognition accuracy (%)			
	Bengali		TIMIT	
	HMMs	DNNs	HMMs	DNNs
Phone-posterior-based tandem PRS	48.97	51.3	62.59	69.7
Consonant-AF-based hybrid PRS	49.37	54.2	61.82	74.2
Vowel-AF-based hybrid PRS	50.96	54.9	63.03	73.8
All-AF-based hybrid PRS	52.17	55.2	63.44	74.4
PP-and-all-AF-based hybrid PRS	52.54	55.8	64.76	74.7

hybrid PRS to develop PP-and-All-AF-based hybrid PRS. In case of *PP-and-All-AF-based hybrid PRS*, $w1$ corresponds to the weighting factor of phone-posterior-based tandem PRS, while $w2$ corresponds to the weighting factor of All-AF-based hybrid PRS.

6.3 Performance Evaluation of Hybrid Phone Recognition Systems

The phone recognition accuracies of the phone-posterior-based and AF-based hybrid PRSs are determined as per the procedure mentioned in Sect. 4.2.3. Table 9 shows the phone recognition accuracies of the phone-posterior-based and AF-based hybrid PRSs. First column lists different types of hybrid PRSs. Second and third columns show the recognition accuracies obtained on Bengali dataset, while the last two columns tabulate the recognition accuracies obtained on TIMIT dataset. It can be observed that the performance of hybrid PRSs is higher than any of the AF-based tandem PRSs (see Tables 6, 8). The improvement in the recognition accuracies of the 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. The All-AF-based hybrid PRSs have higher recognition accuracy compared to phone-posterior-based tandem PRSs. The PP-and-All-AF-based hybrid PRSs developed using DNNs have shown best recognition accuracy on both Bengali and TIMIT datasets. The recognition accuracy of best performing systems is 55.8 and 74.7% for Bengali and TIMIT datasets, respectively.

In all hybrid PRSs, most of the vowels and unaspirated plosives have shown improvement in their classification accuracies, while most of the semivowels, nasals, fricatives and aspirated plosives have shown reduction in their classification accuracies. The reduction in the classification accuracy of the aspirated plosives is mostly due to their misclassification into corresponding unsapirated plosives. The classification accuracies of the vowels are higher in Vowel-AF-based hybrid PRSs compared to Consonant-AF-based hybrid PRSs, while the classification accuracies of the consonants are higher in the Consonant-AF-based hybrid PRSs compared to Vowel AF-based hybrid PRSs. *Silence* has shown improvement in all hybrid PRSs. All-AF-based hybrid PRSs have the classification accuracy of vowels between the classification accuracy of Vowel-AF-based and Consonant-AF-based hybrid PRSs. All-AF-based hybrid PRSs have higher classification accuracy of consonants compared to the Vowel-AF-based and Consonant-AF-based hybrid PRSs. This is mainly due to the improvement in the classification accuracy of the unaspirated plosives. The classification accuracy of semivowels is same in both Consonant-AF-based and All-AF-based hybrid PRSs. PP-and-All-AF-based hybrid PRSs have shown the 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 of all other subgroups. The recognition accuracy of the semivowels in PP-and-All-AF-based hybrid PRS is almost same as that of the baseline PRS.

7 Discussion

In this section, we discuss the performance of the proposed PRSs developed using AFs and compare the results with other state-of-the-art AF-based 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 [44], 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 using AFs to improve the performance of English PRSs, but there are certain difficulties involved in the comparison of the results. Few of these difficulties are listed: (i) Different studies have used different speech corpora with different amount of data (ii) The phones used in development of PRSs are not uniform across all the works, (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 related works and tried to analyze the results. In order to have consistency in comparison across 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 phone recognition accuracies.

In 2001, Kirchhoff et al. have used AFs for conversational speech recognition. They have considered six AF groups—voicing, place, manner, frontness, roundness. The combination of MFCCs and AFs has shown a best recognition accuracy of 72% [22]. Compared to the above work, the performance of the proposed DNN-based PRS (74.7% see Table 9) is much better.

In 2007, Frankel et al. have used AFs for recognition of telephone speech from Fisher and Switchboard corpora. Set of Multilayer Perceptrons (MLPs) are used for predicting the AFs for place, degree (or manner), fricative, nasality, rounding, glottal state, vowel, height, and frontness groups. The combination of Perceptual Linear Prediction Coefficients (PLPCs) and AFs have shown the best recognition accuracies of 40.3 and 37.7% for Fisher and Switchboard corpora, respectively [11]. As an extension of [11], Cetin et al. have shown that use of factored observation model will significantly improve the recognition accuracies of AF-based tandem PRSs. The best results have shown a recognition accuracy of 40.9% (monophone) on Fisher corpora [5]. Although the results of the current work cannot be compared with [11] and [5], due to the difference in the speech corpora used, we feel that the use proposed DNN-based hybrid PRSs on Fisher and Switchboard corpora might give better or at least comparable results with that of [11] and [5].

In 2009, 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 for the context-independent phone recognizer with no LM has a recognition accuracy of 64.84% [43]. For comparing the results of above-mentioned system in [43], we have evaluated the proposed DNN-based PRS with 45 phones, and the recognition accuracy is observed to be 70.20%, which is better compared to above-mentioned system [43].

In 2011, Rasipuram et al. have used the AFs to improve the performance of the PRSs using TIMIT dataset [40]. The AFs are estimated by training two stages of the MLPs. First stage takes PLPCs 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 phone posteriors are used as features to train PRS. The inter-feature dependencies between different AF groups are captured using *multitask learning* approach. The best recognition accuracy reported in [40] is 74.0%, which is less than the performance of the proposed DNN-based hybrid PRS (74.70%).

In 2011, Ghosh et al. have used TVs to improve the phone recognition accuracy. They have considered five broad phone classes, namely vowel, fricative, stops, nasal, and silence for recognition. The best obtained frame-level phone recognition accuracy on development set is 81.28%. The results of current work cannot be directly compared with that of [14], due to the difference in the number of phones and amount of data used. The current work uses 48 phones (for training), while the number of phones considered in [14] is 5 phones. The amount of electromagnetic articulography data considered in [14] is very limited, compared to the TIMIT dataset considered in this study. However, the five phones considered in [14] are quite analogous to the manner AF-predictor considered in this work. From Table 6, the frame-level prediction accuracy of FFNN-based manner AF-predictor using development set is 86.72%, which is better than 81.28%.

In 2014, Mitra et al. have used TVs to improve the performance of continuous speech recognition systems [29]. DNNs are used for estimating eight TVs. DNNs are trained using the TVs and synthetic speech generated for the words of CMU dictionary using Haskins Laboratories Task Dynamic model (TADA). These DNNs are used for predicting the TVs for training and testing set of Aurora-4, the noisy Wall Street Journal (WSJ0) corpus. Aurora-4 contained approximately 15 h duration and

330 test utterances. It was observed that the use of articulatory information in addition to standard cepstral features provides sufficient complementary information that helps to reduce the word error rates. The highest recognition accuracies are obtained using the combination of MFCCs and 30 dimensional principal component analysis-based modulated TVs. The best recognition accuracies obtained for clean speech are 89.1% and 85.4% for matched and mismatched channel conditions, respectively [29]. It is clear that the proposed system has lower recognition accuracy (74.70%) compared to 89.1% reported in [29]. However, it has to be noted that setup in [29] requires us to generate TVs and synthetic speech using TADA for training DNNs to estimate TVs. We need to have a special dataset such as CMU dictionary to accomplish this. But, the proposed setup does not require any special dataset or synthetic speech to predict the AFs. Bi-gram language model with 15 h of data is used in [29], which is better a setup compared to current study. In our future work, we would like to explore combination of TVs and MFCCs described [29] on our database and analyze the performance improvements.

8 Summary and Conclusion

In this work, AFs are explored for developing the phone recognition systems using HMMs and DNNs. The proposed PRSs are evaluated on Bengali and TIMIT speech corpora. The use of articulatory features in addition to the spectral features leads to an improvement in the performance of PRSs. The articulatory features are predicted from the spectral features using FFNNs. 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 the weighted combination approach. The All-AF-based hybrid PRSs outperform the conventional phone-posterior-based tandem PRSs. The All-AF-based hybrid PRSs have higher recognition accuracy compared to the Consonant-AF-based and Vowel-AF-based hybrid PRSs. DNNs have outperformed HMMs in all the cases. The PP-and-All-AF-based hybrid PRSs developed using DNNs have shown best recognition accuracy on both Bengali and TIMIT datasets. The recognition accuracy of best performing systems is 55.8% and 74.7% for Bengali and TIMIT datasets, respectively.

Acknowledgements The work presented in this paper was performed at IIT Kharagpur as a part of the project "*Prosodically guided phonetic engine for searching speech databases in Indian languages*" supported by Department of Information Technology, Government of India.

References

1. A. Afshan, and P. K. Ghosh, Better acoustic normalization in subject independent acoustic-to-articulatory inversion: benefit to recognition. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5395–5399 (2016)
2. E. Armstrong, "IPA Chart : Diphthongs," *Voice and Speech Source*, [Online]. Available : http://www.yorku.ca/earmstro/ipa/diphthongs.html
3. C. S. Blackburn, *Articulatory Methods for Speech Production and Recognition*. PhD Thesis, Trinity College Cambridge and Cambridge University Engineering Department (1996)

4. H.A. Bourlard, N. Morgan, *Connnectionist Speech Recognition: A Hybrid Approach* (Kluwer, Norwell, 1994)
5. O. Cetin, A. Kantor, S. King, C. Bartels, 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)*, pp. 645–648 (2007)
6. 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)
7. S. Dusan and L. Deng, Estimation of articulatory parameters from speech acoustics by Kalman filtering. *Proceedings of CITO Researcher Retreat*, pp. 47–48 (1998)
8. K. Erler, G.H. Freeman, An HMM-based speech recognizer using overlapping articulatory features. J. Acoust. Soc. Am. **100**(4), 2500–2513 (1996)
9. F. Fallside, H. Lucke, T.P. Marsland, P.J. O Shea, M.S.J. Owen, R.W. Prager, A.J. Robinson, and N.H. Russell, "Continuous speech recognition for the TIMIT database using neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 445–448 (1990)
10. J. Frankel, *Linear Dynamic Models for Automatic Speech Recognition*. Ph.D. Thesis, The Centre for Speech Technology Research, University of Edinburgh, UK (2003)
11. J. Frankel, M. Magimai-Doss, S. King, K. Livescu, O. Cetin, Articulatory feature classifiers trained on 2000 hours of telephone speech, in *INTERSPEECH*, pp. 36–41 (2007)
12. J. Frankel, S. King, Speech recognition using linear dynamic models. IEEE Trans. Audio Speech Lang. Process. **15**(1), 246–256 (2007)
13. Gerfen, *Phonetics Theory* [Online]. http://www.unc.edu~gerfen/Ling30Sp2002/phonetics.html
14. P.K. Ghosh, S. Narayanan, Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. Express Lett. **130**(4), 251–257 (2011)
15. A. Graves, A. Mohamed, and G. Hinton, Speech recognition with deep recurrent neural networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6645–6649 (2013)
16. 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)
17. 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)
18. G. John et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, [online]. http://catalog.ldc.upenn.edu/LDC93S1 Linguistic Data Consortium, Philadelphia (1993)
19. H. Ketabdar, H. Bourlard, Enhanced phone posteriors for improving speech recognition systems. IEEE Trans. Audio Speech Lang. Process. **18**(6), 1094–1106 (2010)
20. 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)
21. 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)
22. K. Kirchhoff, G.A. Fink, G. Sagerer, Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. **37**, 303–319 (2002)
23. K. Lee, H. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **37**, 1641–1648 (1989)
24. S. Lee, S. Yildirim, A. Kazemzadeh, S. Narayanan, An articulatory study of emotional speech production, in *INTERSPEECH*, pp. 497–500 (2005)
25. R.P. Lippmann, Neural network classifiers for speech recognition. Lincoln Lab. J. **1**, 107–124 (1988)
26. 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)*, pp. 7145–7149 (2013)
27. V. Mitra, H. Nam, C. Y. Espy-Wilson, Robust speech recognition using articulatory gestures in a Dynamic Bayesian Network framework, in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2011)

28. V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman E, L. Goldstein, Recognizing articulatory gestures from speech for robust speech recognition. J. Acoust. Soc. Am. **131**(3), 2270–2287 (2012)
29. V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, Articulatory features from Deep Neural Networks And Their Role In Speech Recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3017–3021 (2014)
30. P. Mizera and P. Pollak, Improved estimation of articulatory features based on acoustic features with temporal context, in *18th International Conference on Text, Speech, and Dialogue (TSD)*, pp. 560–568 (2015)
31. A. Mohamed, G.E. Dahl, G. Hinton, Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**, 14–22 (2012)
32. S. Narayanan et al., A multimodal real-time MRI articulatory corpus for speech research, in *INTERSPEECH*, pp. 837–840 (2011)
33. S. Narayanan et al., Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). J. Acoust. Soci. Am. **136**(3), 1307–1311 (2014)
34. S.E.G. Ohman, Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soci. Am. **39**(1), 151–168 (1965)
35. J. Pinto, S. Garimella, M. Magimai-Doss, H. Hermansky, H. Bourlard, Analysis of MLP-Based hierarchical phoneme posterior probability estimator. IEEE Trans. Audio Speech Lang. Process. **19**(2), 225–241 (2011)
36. D. Povey et al., The Kaldi speech recognition toolkit, in *IEEE Workshop on ASRU* (2011)
37. V.R. Ramachandran, *Coarticulation Knowledge for a Text-to-Speech System for an Indian Language*, MS Thesis, Speech and Vision Laboratory, Indian Institute of Technology Madras, India
38. L. Rabiner, B. Juang, B. Yegnanarayana, Fundamentals of Speech Recognition. Pearson Education, Upper Saddle River, 2008)
39. M. Rajamanohar, E. Fosler-Lussier, An evaluation of hierarchical articulatory feature detectors, in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 59–64 (2005)
40. R. Rasipuram, M. Magimai-Doss, Improving articulatory feature and phoneme recognition using multitask learning. Artif. Neural Netw. Mach. Learn. **6791**, 299–306 (2011)
41. V.R. Reddy, S. Maity, K.S. Rao, Identification of Indian languages using multi-level spectral and prosodic features. Int. J. Speech Technol. **16**, 489–511 (2013)
42. M. Roch, *IPA/CMU/TIMIT Phone Mappings and American English Examples* [online]. http://roch.sdsu.edu/cs682/IPA-CMU-TIMIT-Phoneset.pdf
43. S.M. Siniscalchi, Chin-Hui Lee, A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition. Speech Commun. **51**, 1139–1153 (2009)
44. 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)
45. The Centre for Speech Technology Research, The University of Edinburgh, MOCHA-TIMIT: MOCHA MultiCHannel Articulatory Database: English, [Online]. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html
46. The International Phonetic Association, *Handbook of the International Phonetic Association*. Cambridge University Press, Cambridge [Online]. http://www.langsci.ucl.ac.uk/ipa/index.html
47. H. Wakita, Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms, in *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 417–427 (1973)
48. S. Wegmann et al., *QuickNet Software and Documentation, Speech Group at International Computer Science Institute* [Online]. http://www.icsi.berkeley.edu/icsi/groups/speech
49. A. A. Wrench, A new resource for production modelling in speech technology. in *Proceedings of the Workshop on Innovations in Speech Processing* (2001)
50. S. Young et al., *The Hidden Markov Model Toolkit and HTK Book*. Cambridge University Engineering Department [Online]. http://htk.eng.cam.ac.uk
51. X. Zhang, J. Trmal, D. Povey, S. Khudanpur, Improving deep neural network acoustic models using generalized maxout networks, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2014)
52. I. Zlokarnik, Adding articulatory features to acoustic features for automatic speech recognition. *J. Acoust. Soc. Am*, **97**, 3246 (1995)