Sādhanā (2020) 45:190
https://doi.org/10.1007/s12046-020-01428-9

Articulatory-feature-based methods for performance improvement of Multilingual Phone Recognition Systems using Indian languages

K E MANJUNATH[1,3,*], DINESH BABU JAYAGOPI[1], K SREENIVASA RAO[2] and V RAMASUBRAMANIAN[1]

[1] International Institute of Information Technology Bangalore, Bengaluru, India
[2] Indian Institute of Technology Kharagpur, Kharagpur, India
[3] U R Rao Satellite Centre, Indian Space Research Organisation, Bengaluru, India
e-mail: ke.manjunath@gmail.com; jdinesh@iiitb.org; ksrao@iitkgp.ac.in; v.ramasubramanian@iiitb.ac.in

MS received 9 May 2019; revised 26 December 2019; accepted 16 April 2020

Abstract. In this work, the performance of Multilingual Phone Recognition System (Multi-PRS) is improved using articulatory features (AFs). Four Indian languages – Kannada, Telugu, Bengali and Odia – are used for developing Multi-PRS. The transcription is derived using *international phonetic alphabets* (IPAs). Multi-PRS is trained using hidden Markov models and the state-of-the-art Deep Neural Networks (DNNs). AFs for five AF groups – place, manner, roundness, frontness and height – are predicted from Mel-frequency cepstral coefficients (MFCCs) using DNNs. The oracle AFs, which are derived from the ground truth IPA transcriptions, are used to set the best performance realizable by the predicted AFs. The performances of predicted and oracle AFs are compared. In addition to the AFs, the phone posteriors are explored to further boost the performance of Multi-PRS. Multi-task learning is explored to improve the prediction accuracy of AFs and thereby reduce the Phone Error Rates (PERs) of Multi-PRSs. Fusion of AFs is done using two approaches: i) lattice re-scoring approach and ii) AFs as tandem features. We show that oracle AFs by feature fusion with MFCCs offer a remarkably low target of PER of 10.4%, which is 24.7% absolute reduction compared with baseline Multi-PRS with MFCCs alone. The best performing system using predicted AFs has shown 3.2% reduction in absolute PER (9.1% reduction in relative PER) compared with baseline Multi-PRS. The best performance is obtained using the tandem approach for fusion of various AFs and phone posteriors.

Keywords. Indian languages; multilingual phone recognition; feature fusion; deep learning; IPA; articulatory features.

1. Introduction

Multilingual Phone Recognition System (Multi-PRS) is a language-independent, universal Phone Recognition System (PRS) that can recognize the phonetic units present in a given speech utterance independent of the language of the speech utterance. The difficulty in developing a Multi-PRS is to arrive at a common multilingual phone-set based on which such a phonetic decoding of input speech utterance can be done independent of its language. In addition to having proper coverage of all the phones occurring across the multiple languages, the common multilingual phone-set should also ensure that the phones of individual languages are accurately mapped to the phones in the common multilingual phone-set. The trancription based on international phonetic alphabets (IPAs) can be used to derive such a common multilingual phone-set, which involves mapping the acoustically similar phonetic units across languages to an underlying IPA unit. Since the IPAs have strict one-to-one correspondence between symbols and sounds, the transcription of all the distinct languages of the world can be derived using IPAs [1].

One of the important direction in multilingual speech recognition is the use of articulatory features (AFs), given that their production basis serves as a common feature set across languages. AFs are related to the speech production mechanisms and provide a more compact representation of the speech independent of the language in which it is produced. AFs represent the positioning and movements of articulators during the production of a sound unit. The speech production involves the articulators such as glottis, velum, tongue, hard palate, teeth, lips and alveolar ridge.

The AFs represent a higher degree of *invariance* and hence it is more appropriate to use them in multilingual tasks that will represent the acoustic-phonetic variability across languages [2]. The IPA chart is designed based on

[*] For correspondence

Published online: 30 July 2020

the speech production characteristics (i.e. AFs representation) of each sound unit [1]. Each sound unit has unique AFs and they change from one sound unit to another. Five broad AF groups – height, frontness, roundness, manner and place – are considered in this study. The consonants characteristics are captured by manner and place AFs, while the vowels characteristics are captured by height, roundness and frontness AF groups. The significance of having five AF groups to capture various AFs is reported in [3, 4].

Three ways of deriving AFs are mainly explored in literature: i) direct physical measurements [5–10], ii) classification scores for pseudo-articulatory features [11–14] and iii) acoustic-articulatory transformations using inverse mapping [15–17]. Among the three approaches, the second approach is more feasible and most widely used in the context of continuous speech recognition compared with other two approaches. The first method is difficult to explore due to the difficulty of getting speech corpora having the articulatory motions measured using physical instruments for Indian languages, while the third approach suffers as there are no inverse mapping techniques available for deriving all the AFs. Hence, we have explored the second approach in this study.

In this work, our focus is on improving the performance of Multi-PRS using AFs. We examine Deep Neural Network (DNN)-based AF estimation from Mel-frequency cepstral coefficients (MFCCs), and use an early-fusion framework to augment the MFCC feature vector with various categories of AFs to enhance the multilingual phone recognition performance. We also examine the use of multi-task learning (MTL) to improve the AF prediction accuracies and thereby contribute to the improvement in the performance of Mutli-PRSs. In essence, this work focuses on how best to arrive at a feature space (the AF parameter space) and the common multilingual phone-set (the IPA set) that ensures enhanced *invariance* of the phonetic units amidst the increased variability due to the multilingual nature of the Multi-PRS problem.

The rest of the paper is organized as follows. Section 2 provides the literature survey on related works. Section 3 describes our experimental set-up. Detailed description of development of Multi-PRS is provided in section 4. Section 5 describes the extraction of AFs. The use of AFs and the feature fusion techniques are given in section 6. The use of MTL to enhance AF-predictors and Multi-PRSs is described in section 7. Section 8 provides the summary of the paper.

2. Related work

Some of the notable works related to multilingual speech recognition reported in the literature are as follows. In 1998 Corredor-Ardoy *et al* [18] explored the development of a Multi-PRS using spontaneous telephone speech of four languages, namely British English, French, Castillan Spanish and German. In 2001, Schultz *et al* developed a multilingual speech recognition system using GlobalPhone database in a language-independent, -dependent and language-adaptive manner. The multilingual acoustic models are used to estimate the acoustic models for a new language in a fast and efficient way [19–21]. In 2013, Heigold *et al* [22] trained multilingual acoustic models using DNNs and compared them to monolingual and cross-lingual systems. Data from 11 Romance languages with a total amount of 10k hours was used for conducting experiments. It is found that multilingual systems outperform both cross-lingual and monolingual systems [22]. In 2014, Vu *et al* [23] developed multilingual DNN-based acoustic modelling that can be applied to new languages. The effect of phone merging on multilingual DNN in the context of rapid language adaptation is investigated. Ten different languages from the GlobalPhone database are considered [23].

Although there have been significant efforts in developing multilingual speech recognizers, the number of works exploring the development of multilingual speech recognizers in the context of Indian languages is very limited. A few works in this direction are as follows. In 2005, Kumar *et al* [24] developed a Hidden Markov Model (HMM)-based bilingual speech recognizer using Tamil and Hindi languages. The Bhattacharyya distance measure is used to group the acoustically similar phones across two languages [24]. In 2005, Gangashetty *et al* [25] developed a multilingual speech recognition system based on syllable units using 3 Indian languages – Telugu, Hindi and Tamil. The syllable-like consonant–vowel units across 3 languages are merged to train the multilingual speech recognizer [25]. In 2014, Mohan *et al* [26] developed a small vocabulary multilingual speech recognizer using two linguistically similar Indian languages – Hindi and Marathi. It can be noted that none of the multilingual efforts has examined the use of IPA to derive a common phone-set labelling mechanism in the context of Indian languages and all are limited to simplistic approaches. Our work, based on the IPA transcription, represents an unifying framework generalizable to new languages easily.

Since the AFs are more universal [27–29] and less dependent on language compared with the conventional spectral features [30–32], they can be explored to improve the performance of multilingual speech recognizers. The AFs can be continuous or discrete [33], with the Mermelstein model [34–36] being a classic example of the continuous model. AFs have been consistently shown to improve the performance of speech recognizers, such as in [37–39] (using continuous valued AFs) and in [11–13] (using discrete valued AFs). Although the AFs are widely used to improve the performance of monolingual speech recognizers [11–13, 40, 41], only a few works exploring the AFs to improve multilingual speech recognizers are reported. The number of works exploring the use of AFs to improve the performance of multilingual speech recognizers in the context of Indian languages is very limited. A few notable works in this direction are as follows.

In 1997, Deng [27] proposed a integrated-multilingual speech recognizer framework mainly focusing on cross-language portability. The articulatory, acoustic and auditory features are used for capturing the cross-language commonality. The AFs are derived from the dynamic properties of the vocal tract that are derived using the task-dynamic model. The tract variables such as upper and lower lips, jaw, tongue body, tongue tip, velum, glottal width, total lung force, supralaryngeal vocal tract volume and vocal fold tension are considered for representing the AFs. The study aims to a build a universal speech recognizer that can be used across all the languages [27].

In 2003, Stuker *et al* [42] showed that the AFs derived from cross-lingual and multilingual AF detectors can reduce the Word Error Rates (WERs) of HMM-based speech recognizers significantly. The AF detectors can compensate the inter-language variability. The study considers five languages – Chinese Mandarin, German, Japanese, Spanish and English – whose data is taken from GlobalPhone [43] and Wall Street Journal corpora. The transcription was derived using IPA symbols. It is found that the feature detectors that are trained using the multilingual AFs (i.e. the AFs of multiple languages) have higher classification accuracy compared with the feature detectors trained using the AFs of single language. The performance of speech recognizers based on multilingual AF detectors is superior compared with the speech recognizers based on monolingual AF detectors. The use of multilingual AFs has significantly reduced the WER of HMM-based recognizer [2, 42].

In 2007, Ore [44] developed AF detectors for detecting the AFs using Gaussian Mixture Models (GMMs) and Multi-Layer Perceptrons (MLPs). English dataset from Wall Street Journal corpus and German, Spanish and Japanese datasets from GlobalPhone corpora [43] are used. Multilingual AF detectors are developed using the data from all four languages. Four monolingual AF detectors separately for each language are also developed. The outputs of the AF detectors were used as features for training HMM-based phoneme recognizer. It is shown that the AFs output by the multilingual AF detectors performs better than that of the monolingual AF detectors. It is also found that the speech recognizers using AFs have higher performance compared with MFCCs [44].

In 2011, Rasipuram and Magimai-Doss [45] used the MTL to improve the prediction accuracies of MLP-based AF estimators. It is shown that the use of MTL-derived AFs has significant improvement in the performance of TIMIT phoneme recognizer [45].

In 2016, Muller *et al* [46] demonstrated the development of speech recognizers for low-resource languages using multilingual speech corpora. AFs are used to improve the performance of multilingual speech recognizers. Fully connected feed-forward neural networks are used for predicting the AFs. The datasets of 4languages – English, French, German and Turkish (taken from *Euronews* corpus) – are considered for training the multilingual speech recognizer using DNNs. Multilingual phone-set is derived by merging the IPA symbols from all the languages. The combination of AFs and *language feature vectors* has shown the least WER [46]. The use of *language feature vectors* to improve the performance of multilingual systems is described in [47].

In 2017, Sahraeian [48] worked on the adaptation of multilingual DNNs to a low-resource language. Articulatory-like features are extracted using a feature transformation technique called Intrinsic Spectral Analysis (ISA) manifold learning. Data of 9 different languages from GlobalPhone corpus [43] are considered in their studies. The language independence behaviour of spectral and ISA features is studied using 7 AFs – front vowel, back vowel, open vowel, plosive, labial, nasal and fricative – which are extracted from the bottleneck layer of DNNs. It is found that the ISA features exhibit overall better language-independent behaviour than spectral features. Several experiments were conducted under monolingual, cross-lingual and multilingual settings to demonstrate the usefulness of ISA. It is shown that the ISA features have significant advantages compared with traditional filter-bank features in multilingual and low-resource scenarios [32, 48, 49].

In 2018, Dash *et al* [50] explored the use of articulatory information to improve the automatic speech recognizer (ASR) performance using 4 Indian languages, namely Hindi, Marathi, Bengali and Oriya. Articulatory movements were recorded during speech production using an electromagnetic articulograph and trained together with acoustic features to build ASRs for these languages. ASR systems are trained using MM-HMM, DNN-HMM and Long Short Term Memory recurrent neural network (LSTM)-HMM. A multilingual, multi-modal speech recognizer that was built by constructing a unified dictionary consisting of common and unique phonemes of all the four languages has shown significant reductions in the phoneme error rates [50].

The objective of our study is to examine the use of AFs to improve the performance of Multi-PRSs. This study is perhaps the first of its effort in the context of Indian languages in several fronts, such as the use of IPA-based transcription to derive the common multilingual phone-set for Multi-PRS, the application of MTL for estimation of AFs, the use of DNN-derived AFs as features with improved Phone Error Rate (PER) and establishing very low PERs ($\sim 10\%$) for *oracle AFs*, thereby setting the baseline performance achievable if AFs can be estimated accurately from speech directly or via other spectral representations. Earlier, we have examined the use of AFs to improve the performance of monolingual PRSs using Bengali and TIMIT datasets [3]. The work presented here is an extension of the work reported in [51].

3. Experimental set-up

The following subsections provide a detailed description of the experimental set-up used in this study.

3.1 *Multilingual speech corpora*

The multilingual speech corpora are developed using 4 Indian languages – Telugu (TE), Kannada (KN), Odia (OD) and Bengali (BN). The speech corpora were developed as a consortium project titled *Prosodically guided phonetic engine for searching speech databases in Indian languages* supported by DIT, Government of India [52]. Further details on speech corpora can be found in [53–57].

A sampling rate of 16 kHz with a precision of 16 bits per sample is used for recording the audio files. The IPA chart was used for deriving the phonetically rich transcription for all the wave files. Independent of the language of production of a sound unit, an IPA will always have same production characteristics and hence IPA is language independent in nature. Since the IPAs can be used to transcribe sound units of any language, the merging of acoustically similar phonetic units from multiple languages based on IPA transcription will be more accurate compared with the transcription based on ascii text.

The speech corpora in reading mode, collected from the television and radio news broadcasts and the reading of textbooks and story-books in a closed room noiseless environment, is considered in this work [53, 57]. Each audio file contains 1 sentence of speech utterance. The training and testing consists of non-overlapping speakers with a split ratio of 80 : 20 for train and test dataset, respectively.

Table 1 shows the various statistics of multilingual speech corpora used in this study. The duration of speech data and the count of speakers are separately shown for each language. The language name is listed in the first column. The count of male and female speakers is provided in the next 2 columns. Fourth to seventh columns tabulate the duration of different datasets in terms of number of hours.

3.2 *Training HMMs and DNNs*

We start our training by building Context-Independent (CI) GMM-HMMs (referred to as HMMs throughout) using flat-start initialization. The training of Context-Dependent (CD) HMMs is initialized using the alignments obtained from the CI HMMs. Further, the CD DNN-HMMs (referred to as DNNs throughout) are trained using the alignments obtained from the CD HMMs. The training of CI DNNs is also explored using the alignments generated by the CI HMMs. Monophones are used for training CI models, while the triphones are used for training CD models.

Table 1. Statistics of multilingual speech corpora.

	#Spkrs		Duration (h)			
Language	Male	Female	Train	Dev	Test	Total
Telugu	9	10	4.05	0.47	1.07	5.59
Kannada	7	9	2.80	0.33	0.76	3.89
Odia	14	16	3.58	0.36	0.97	4.91
Bengali	20	30	3.42	0.40	0.99	4.81

The acoustic-phonetic decision tree is used for capturing the mapping from HMM-state index and the phonetic context, to an emission probability density [58]. Number of transition states, number of Gaussians and number of transition IDs depend on the context being modelled and the number of phones used. DNNs having the hidden layers with tanh non-linearity and the output layer with softmax activation are considered. Greedy layer-by-layer supervised training is employed for training DNNs. A learning rate of 0.015 was used initially, which was then exponentially decreased for first 15 epochs. Last 5 epochs use a constant learning rate of 0.002. After the completion of addition of all the hidden layers to the network, the parameters of each layer are scaled separately by performing shrinking after every 3 iterations. Halfway between the end of training and completion of addition of all the hidden layers, mixing up was carried out.

The preconditioned affine components are used for maintaining the stability of training. After the completion of final iteration of DNN training, a single model is obtained by combining the models of last 10 iterations. A temporal context of 9 frames, containing 4 frames on both the sides, is used as input to DNNs. The optimal numbers of hidden layers for Multi-PRSs and AF-predictors are tuned, by varying the width of hidden layers. We found that the 4 hidden layered DNNs are suitable for AF-predictors (see section 5.2), and the 5 hidden layered DNNs are good for Multi-PRSs (see section 4). For example, the baseline Multi-PRS described in section 4 has 432, 300, 19860 units at input, hidden and output layers, respectively. Depending on the dimension of the input features, the total number of parameters of DNNs ranges between 1.9 and 2.0 millions.

Decoding is done using bi-phone (phoneme bi-grams) language model. The lattices are decoded using the acoustic scaling factor and language model weighting factor, which are optimally determined from the development set so as to minimize the PER. The procedure used for training the DNNs is similar to the one described in [59]. DNNs training used in this study is similar to the one presented in [59]. The open-source *Kaldi* toolkit is used for building the phone recognition models [60].

4. Development of monolingual and multilingual PRSs

Monolingual Phone Recognition Systems (Mono-PRSs) are trained using the data of single language, while the Multi-PRSs are trained using the data from multiple languages. In this work, we have considered 4 Indian languages – TE, KN, OD and BN. The data from the multilingual speech corpora described in section 3.1 is considered. The *train* data of all the four languages shown in table 1 is used for training the Multi-PRS. Similarly, the *test* data of all the four languages shown in table 1 becomes the test data for evaluating the performance of Multi-PRS. For each language, a separate Mono-PRS is developed using the corresponding *train* and *test* data shown in table 1. This results in development of 4 Mono-PRSs for TE, KN, OD and BN languages. The common phone-set for each Mono-PRS is derived by grouping the acoustically similar IPAs present in the training data of the corresponding language and selecting the phonetic units that have sufficient number of occurrences to train a separate model for each of them. The IPAs that do not have sufficient number of occurrences will be mapped to the closest linguistically similar phonetic units present in the considered common phone-set. The count of phones present in the common phone-sets of TE, KN, OD and BN languages is found to be 35, 36, 36 and 34, respectively. In case of Multi-PRS the common multilingual phone-set is determined using the following procedure: The acoustically similar IPAs from all the four languages are merged together, and the phonetic units having sufficient number of occurrences (so as to train a separate model for each of them) are added to common multilingual phone-set. IPAs with insufficient number of occurrences are mapped to the closest linguistically similar phonetic units present in common multilingual phone-set. The count of phones present in the common multilingual phone-set of Multi-PRS is found to be 44. MFCC features are used for building Mono-PRSs and baseline Multi-PRS. The procedure for extracting the MFCCs is similar to the one described in [61]. We have explored both DNNs and HMMs for training the phone recognizers under CD and CI settings. The procedure given in section 3.2 is used for training DNNs and HMMs.

We have used the sclite tool [62] for computing PERs. The hypothesized text decoded by the speech recognizer is compared to the reference transcription using Dynamic Programming (DP) to compute PER. The cost of insertions (I), correct phones, substitutions (S) and deletions (D) in DP string alignment is 3, 0, 4 and 3, respectively. The PER is computed using Equation (1):

$$Phone\ Error\ Rate = \frac{S + D + I}{N} \times 100\% \tag{1}$$

where N indicates the total number of phones in the reference transcriptions.

Table 2 shows the PERs of Mono-PRSs and baseline Multi-PRS. PER is computed by comparing the decoded phones to the reference phone labels. In table 2, as one moves from right to left, PERs increase in all the rows. This shows that the CI models have higher PERs than CD models. DNNs have lower PERs (better performance) compared with HMMs. Since the CD DNNs have shown the least PERs in all the cases, we have used only CD DNNs in all our further experiments. Detailed description on development of Multi-PRS can be found in [61, 63].

Table 2. PERs of monolingual and baseline Multilingual Phone Recognition Systems developed using MFCCs.

	CI		CD	
PRS	HMM	DNN	HMM	DNN
Telugu	42.1	35.5	35.0	30.7
Kannada	43.5	39.5	38.5	37.1
Odia	33.6	29.5	28.0	26.5
Bengali	49.0	41.6	43.4	37.6
Multi-PRS	49.4	39.8	39.0	*35.1*

5. Extraction of AFs

The prediction of the AFs from the spectral features using DNNs is discussed in detail in the following subsections.

5.1 *AFs*

Each sound unit can be represented as a set of features based on the articulators used to produce it. These features, which describe the properties of speech production of a sound unit, are called AFs. In addition to providing the discriminating features between various phonetic units, the AFs will also enable us to capture the co-articulation effect between the neighbouring phonetic units [37, 64–66]. The AF specification containing the AF values for each of the five AF groups – height, frontness, roundness, manner and place – is shown in table 3. The AF group is shown in the first column. The possible feature values (i.e. AF specification) for each AF group are shown in the second column. The last column shows the cardinality. The number of feature classes in an AF group is indicated by cardinality. Similar kinds of AF specifications are used in [11–13].

5.2 *Prediction of AFs using AF-predictors*

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. AFs are predicted for five AF groups – height, frontness, roundness, manner and place – using AF-predictors. MFCCs are used as features for training the AF-predictors using DNNs. The procedure for extracting MFCCs is similar to the one described in [61]. For training DNNs we require the speech data, which is transcribed using AF labels at frame-level. Since the transcription is available at phone-level, the frame-level AF labels for each AF group

Table 3. Articulatory feature specification for different AF groups of multilingual speech corpora.

AF group	Features	Cardinality
Place	silence, vowel, glottal, velar, palatal, retroflex, alveolar, labiodental, bilabial	9
Manner	silence, vowel, nasal, approximant, fricative, plosive	6
Roundness	silence, consonant, unrounded, rounded	4
Frontness	silence, consonant, back, mid, front	5
Height	silence, consonant, open, open-mid, close-mid, close	6

are obtained by mapping the phone labels present in the phone-level transcription to AF label. 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 3. Table 4 shows the mapping of each phone label into a set of AF labels of various AF groups. The first column in table 4 lists unique IPA symbols present in the IPA transcription of multilingual speech corpora. The second to sixth columns show the corresponding place, manner, roundness, frontness and height AF values for each phone. The mapping for each IPA symbol to various AF groups is derived using the IPA chart [1].

AF-predictors are trained for classification of the features shown in table 3. The posterior probabilities generated by the AF-predictors represent AFs. Figure 1 illustrates the prediction of manner AFs (i.e. manner AF-predictor). The predicted feature values represent the manner AFs. The block diagrams for remaining four AF-predictors are similar to figure 1.

Figure 2 illustrates the development of AF-predictors for five AF groups. The AFs for a particular AF group are predicted using the AF-predictor of that specific group.

5.3 *Performance evaluation of AF-predictors*

The performance of AF-predictors is evaluated using 3 measures: i) frame-wise accuracy, ii) mean squared error (MSE) and iii) AF-Estimation Error Rate (AF-EER). The frame-wise accuracy of each AF-predictor is computed by comparing the decoded AF label to the actual AF label at frame-level [12, 13]. MSE measures the average of the squares of the errors between the predicted and oracle AFs. AF-EER of AF-predictors is computed by comparing the decoded AF labels to the reference AF labels using DP. The procedure is similar to the computation of PER [62]. AF-EER is computed similar to the computation of PER using Equation (1) as described in section 4 except that the AF labels are used for comparison in place of the phone labels.

Table 5 shows the frame-wise accuracy, MSE and AF-EER of various AF-predictors. The first column shows AF group. The second to fourth columns show the corresponding frame-wise accuracy, MSE and AF-EER. *Roundness* AF group shows the highest frame-wise accuracy, while the *height* AF group shows the least frame-wise accuracy. *Place* AF group shows the least MSE, while the *height* AF group shows the highest MSE. *Manner* AF group shows the least AF-EER, while the *height* AF group shows the highest AF-EER.

6. AFs for multilingual phone recognition

The predicted AFs and MFCCs are combined to improve the performance of Multi-PRSs. The combination of predicted AFs and MFCCs is carried out using 2 approaches, namely i) lattice re-scoring approach (LRA) and ii) combining AFs as tandem features (AF-Tandem). In addition to AFs, we have also explored the combination of phone posteriors (PPs) to further boost the performance of Multi-PRS. Figure 3 shows a block diagram of combination of AFs using LRA. There are 3 stages in figure 3. In the first stage, the AF-predictors are developed to predict the AFs for 5 AF groups from MFCCs. DNNs are used to develop AF-predictors. In the second stage, the predicted AFs (output of first stage) are combined with the MFCCs to develop Multi-PRSs. Since these Multi-PRSs are developed using AFs and are arranged in tandem, we call them AF-based tandem Multi-PRSs. The third stage is developed to combine the AFs from multiple AF groups. In the third stage, LRA is used for combining the AF-based tandem Multi-PRSs developed in the second stage.

In the AF-Tandem approach of combining the AFs, predicted AFs from different AF-predictors are augmented with MFCCs and used as tandem features to develop Multi-PRSs. Figure 4 shows a block diagram of proposed Multi-PRS that uses AF-Tandem approach for combining the AFs of different AF groups.

6.1 *Development of AF-based tandem multi-PRSs*

We have developed AF-based tandem Multi-PRSs using the *combination of MFCCs and the AFs*. The training data used for development of AF-based tandem Multi-PRSs is the same as the training data of Multi-PRS described in section 4, except that the *combination of MFCCs and AFs* is used as features instead of MFCCs alone. The AFs for

Table 4. Mapping of phone labels in multilingual speech corpora to AF values of various AF groups.

Phones	Articulatory feature groups				
	Place	Manner	Roundness	Frontness	Height
a æ	vowel	vowel	unrounded	front	open
ɛ	vowel	vowel	unrounded	front	open-mid
e	vowel	vowel	unrounded	front	close-mid
i ɪ	vowel	vowel	unrounded	front	close
ɑ	vowel	vowel	unrounded	back	open
u ʊ	vowel	vowel	rounded	back	close
ɔ	vowel	vowel	rounded	back	open-mid
ɐ	vowel	vowel	unrounded	mid	open
ə	vowel	vowel	unrounded	mid	open-mid
o	vowel	vowel	rounded	back	close-mid
ɘ	vowel	vowel	unrounded	mid	close-mid
ɒ	vowel	vowel	rounded	back	open
œ	vowel	vowel	rounded	front	open-mid
k k^h ɡ ɡ^h q	velar	plosive	consonant	consonant	consonant
p p^h b b^h	bilabial	plosive	consonant	consonant	consonant
t t^h d d^h	alveolar	plosive	consonant	consonant	consonant
ʧʧ^h ʤʤ^h	palatal	plosive	consonant	consonant	consonant
ʈ ʈ^h ɖ ɖ^h	reftroflex	plosive	consonant	consonant	consonant
ɾ ɹ r l	alveolar	approximant	consonant	consonant	consonant
v ʋ	labiodental	approximant	consonant	consonant	consonant
j	palatal	approximant	consonant	consonant	consonant
ɭ	reftroflex	approximant	consonant	consonant	consonant
m	bilabial	nasal	consonant	consonant	consonant
ŋ	velar	nasal	consonant	consonant	consonant
n	alveolar	nasal	consonant	consonant	consonant
ɲ	palatal	nasal	consonant	consonant	consonant
n	reftroflex	nasal	consonant	consonant	consonant
s s ʃ ʒ z	alveolar	fricative	consonant	consonant	consonant
h	glottal	fricative	consonant	consonant	consonant
f	labiodental	fricative	consonant	consonant	consonant
x	velar	fricative	consonant	consonant	consonant
sil	silence	silence	silence	silence	silence

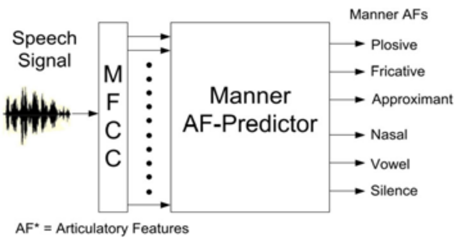

Figure 1. Block diagram of manner articulatory features predictor.

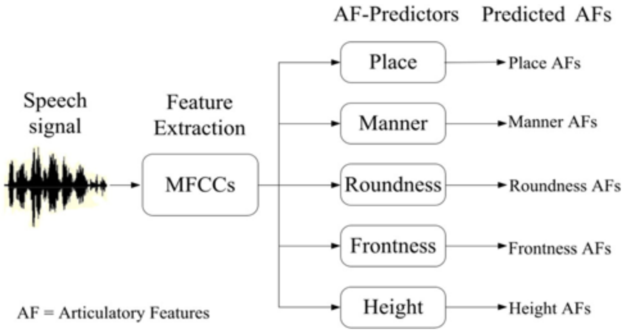

Figure 2. Block diagram for the prediction of articulatory features.

Table 5. Frame-wise accuracies, mean squared errors and AF-Estimation Error Rates of various AF-predictors.

AF group	Frame-wise accuracy (%)	Mean squared error	AF-estimation error rate (%)
Place	85.6	0.025	21.3
Manner	89.4	0.028	17.9
Roundness	90.8	0.037	18.5
Frontness	84.8	0.048	23.3
Height	80.5	0.051	26

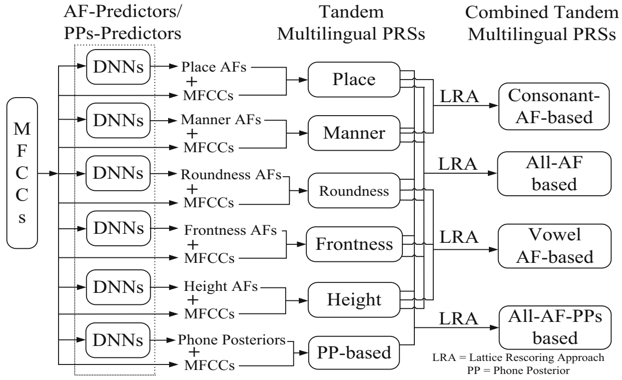

Figure 3. Block diagram of proposed Multi-PRS using lattice re-scoring approach for fusion of articulatory features.

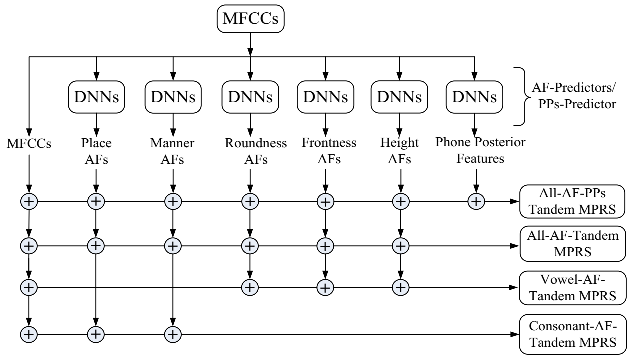

Figure 4. Block diagram of proposed Multi-PRS using AF-Tandem approach for fusion of articulatory features.

each AF group are predicted from the spectral features using the AF-predictors, as per the procedure mentioned in section 5. In the tandem approach, DNNs are first trained using MFCCs to perform the classification at frame-level and then the frame-level posterior probability estimates of the DNNs are used as features for developing Multi-PRSs. The predicted AFs of a particular AF group are augmented with the MFCCs to develop AF-based tandem Multi-PRS for that AF group [67, 68]. Separate tandem Multi-PRSs are developed using the AFs predicted from each AF group. This leads to development of 5 different AF-based tandem Multi-PRSs.

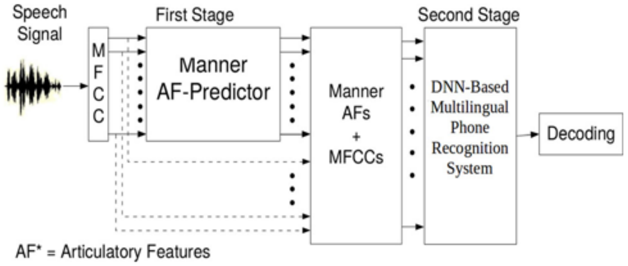

Figure 5. Block diagram of the manner-AF-based tandem Multi-PRS.

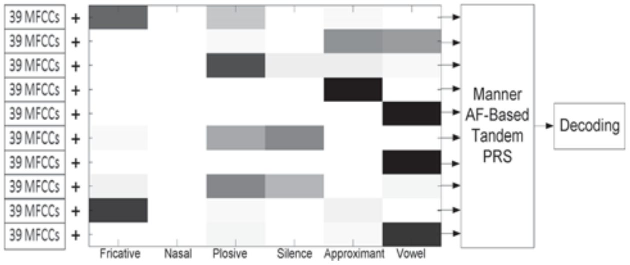

Figure 6. Illustration of manner-AF-based tandem PRS for 10 frames using posteriogram representation.

Figure 5 shows a block diagram of manner-AF-based tandem Multi-PRS. The manner AF-predictor is used for predicting the manner AFs, as shown in figure 1. The combination of predicted manner AFs and MFCCs are used as features for training DNNs to develop a manner-AF-based tandem Multi-PRS in the second stage. Similarly, five different AF-based tandem Multi-PRSs are developed using the predicted AFs from each AF group.

Figure 6 illustrates the manner-AF-based tandem PRS for 10 frames using posteriogram representation. The MFCCs are augmented with the posteriogram distribution of manner AFs obtained in the first stage. The combination of the MFCCs and the manner AFs is then fed to the manner-AF-based tandem Multi-PRS for decoding the phones in the input speech utterance.

Five AF-based tandem Multi-PRSs are developed using the combination of MFCCs and the AFs predicted by the corresponding AF-predictor. To establish the target performance achievable by the predicted AFs, the oracle AFs for each AF group are obtained as follows. The phone labels are mapped to AF labels at frame level. The frame-level posteriogram for oracle AFs is generated by setting the posterior corresponding to the AF label to 1 and remaining posteriors to 0. The posteriogram thus generated will be used as oracle AFs.

Table 6 shows the PERs of AF-based tandem Multi-PRSs. The results are shown separately for predicted and oracle AFs. The second column shows the PERs of AF-based tandem Multi-PRS using *combination of MFCCs and predicted AFs* as features, while the third column shows the PERs obtained using *combination of MFCCs and oracle AFs* as features. For better analysis and comparison of results, the performance of baseline Multi-PRS using MFCCs is also shown in the table. It is observed that the PERs of all the tandem Multi-PRSs are superior compared with the baseline Multi-PRS. This clearly indicates that the use of AFs has reduced the PERs. The average PER of oracle AFs is 14.5% lower than that of predicted AFs. This indicates that there is large scope to reduce the PERs of predicted AFs (up to 14.5% on an average). In addition to the proposed DNN-based predicted AFs, alternative methods for predicting the AFs can be explored including continuous valued AFs.

The *place*-AF-based tandem Multi-PRS shows the highest reduction in PER, and *roundness*-AF-based tandem system shows the least reduction using predicted AFs. This is because *place* AF group has the highest cardinality (i.e. 9), while the *roundness* has the least cardinality (i.e. 4) as shown in table 3. The cardinality indicates number of feature classes (i.e. feature dimension). Higher cardinality (higher feature dimension) provides more discriminative information to classify among various phonetic units. This results in improved phone recognition accuracy and reduces the PER. Similarly, lower cardinality would lead to higher PER. The consonant AF-based systems have lower PERs compared with vowel AF-based systems. It is found that misclassifications among the consonants are reduced in consonant AF-based systems, and the misclassifications among the vowels are reduced in vowel-AF-based systems.

Table 6. PERs of AF-based tandem Multilingual Phone Recognition Systems.

	PER (%) of CD DNNs	
Features	Predicted AFs	Oracle AFs
MFCCs (baseline)	35.1	35.1
MFCCs + place	33.5	21.1
MFCCs + manner	34.1	24.0
MFCCs + round	34.9	26.8
MFCCs + front	34.1	26.9
MFCCs + height	34.3	23.1

6.2 *Combination of AFs from multilple AF groups*

The AFs from different AF groups are combined together to take the mutual advantage of all the AFs at the same time. We have explored 2 approaches for combination: i) LRA approach and ii) AF-Tandem approach. In the LRA approach, the lattices generated by the AF-based tandem systems are combined using the lattice re-scoring method [69]. The weighting factors required for LRA are tuned using development set. In the AF-Tandem method of combination, AFs are augmented as tandem features along with MFCCs to develop Multi-PRSs [12, 13]. The AFs derived from the consonant AF groups are combined to develop consonant-AF-based Multi-PRS, while the vowel-AF-based Multi-PRS is developed by combining the AFs from vowel AF groups. All-AF-based Multi-PRS is developed by combining all the 5 AF-based tandem systems.

Further, we have also explored combining the PPs along with all the predicted AFs to develop All-AF-PP-based Multi-PRS [70]. Similar to AFs, the PPs are predicted from the MFCCs by training a DNN [3, 68, 71, 72]. The posterior probabilities of different phones in each frame are given by $p(q_t = i|x_t)$, where q_t is a phone at time t, $i = 1, 2, \ldots, N$ and x_t is the acoustic features at time t such that

$$\sum_{i=1}^{N} P(i) = 1, \tag{2}$$

where N is the total number of phone classes.

Table 7 shows the PERs of different AF-based Multi-PRSs combined using LRA and AF-Tandem approaches. The results are shown separately for predicted and oracle AFs. The improvements in the performance are consistent. The consonant-AF-based has higher PER reduction compared with vowel-AF-based, while the All-AF-based has higher PER reduction compared with consonant-AF-based system. The PER of All-AF-based Multi-PRS using oracle AFs is 22.3% lower than that of predicted AFs. Given the remarkably low PER of $\sim 10\%$ for oracle-based Multi-PRS, there is much scope for enhanced prediction of AFs to improve the Multi-PRS to reach the performance of oracle AFs.

It is observed that the LRA method of combination has least PERs for consonant-AF-based, vowel-AF-based and All-AF-based Multi-PRSs, while the AF-Tandem method of combination has shown least PERs for All-AF-PP-based Multi-PRS. Since the oracle PPs are same as the ground truth reference labels, it does not make any sense to use oracle PPs as the features. Hence, we have not conducted any experiments related to All-AF-PP-based Multi-PRS using oracle PPs and the corresponding values in table 7 are represented as '-' (hyphen) indicating *not applicable*.

The AF-Tandem method (through All-AF-PP-based Multi-PRS) shows the least PER of 32.3% with an absolute reduction of 2.8% in the PER (8% reduction in relative PER) compared with baseline Multi-PRS. The AF-Tandem

Table 7. PERs of combined tandem Multilingual Phone Recognition Systems.

	Predicted AFs		Oracle AFs	
Combined Multi-PRSs	LRA	AF-Tandem	LRA	AF-Tandem
Vowel-AF-based	33.4	34.8	22.1	21.8
Consonant-AF-based	33.0	33.7	19.6	17.8
All-AF-based	32.7	33.5	12.9	10.4
All-AF-PP-based Multi-PRS	32.6	*32.3*	–	–

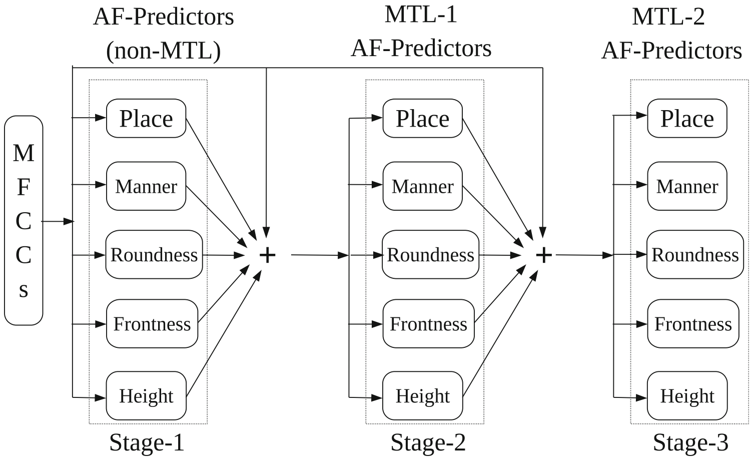

Figure 7. Block diagram of the development of AF-predictors using multi-task learning.

Table 8. AF-Estimation Error Rates of various AF-predictors used with and without MTL approaches.

	AF-Estimation error rate (%)		
AF-predictor	Non-MTL (stage-1)	MTL-1 (stage-2)	MTL-2 (stage-3)
Place	21.3	19.6	19.5
Manner	17.9	16.7	16.5
Roundness	18.5	16.4	16.0
Frontness	23.3	20.6	20.3
Height	26	22.8	22.3

Table 9. Phone Error Rates of combined Multi-PRSs using both MTL- and non-MTL-based AF-predictors.

	Phone error rate (%)		
Combined Multi-PRSs	Non-MTL	MTL-1	MTL-2
All-AF-based	33.5	33.6	33.6
All-AF-PP-based	32.3	32.4	*31.9*

method not only performs better than LRA but also has less complex structure than that of LRA. The time complexity of LRA is almost 5× higher than that of AF-Tandem in terms of both training and decoding.

There are 33 consonants and 11 vowels in the phone-set considered. Around 55% of the test data is made of consonants, whereas only 45% constitutes vowels. Out of 45% of vowel data 15% is wrongly classified, while 26% out of 55% of consonant data is wrongly classified. This means that there is a larger scope to reduce the misclassifications within the consonants than vowels. Since the consonant AFs mainly reduce the misclassifications within the consonants and there is larger scope to reduce the misclassifications within the consonants, the consonant-AF-based Multi-PRS has shown higher improvement in PERs compared with the vowel-AF-based Multi-PRS. Since there are only a few vowel classes the vowels classification using MFCCs itself provides a reasonably good recognition accuracy and there is not much scope for further improvement in the recognition accuracies using vowel AFs, which reduce the misclassifications among vowels. Also, the number of discriminative feature classes in consonants AFs is higher than that of vowel AFs.

7. MTL for performance enhancement

The theory of MTL states that by jointly learning different related tasks that share the same input and some internal representation, the performance of each task can be improved [48, 73]. In the following subsections, we investigate the use of MTL to improve the performance of AF-predictors and Multi-PRSs.

7.1 *Development of AF-predictors using MTL*

In this section, we investigate the MTL approach for joint estimation of AFs of various AF groups. The MTL approach is explored to further reduce the PERs of AF-predictors described in section 5. The approach used for development of MTL-based AF-predictors is similar to the one described in [45]. Figure 7 shows a block diagram of the development of AF-predictors using MTL approach. The block diagram has 3 stages. Stage-1 is the same as the one shown in figure 2. In *stage-1*, DNNs are trained to develop AF-predictors using MFCCs as features. These AF-predictors are not based on MTL and are called *non-MTL AF-predictors*. In *Stage-2*, MFCCs and AFs from all the 5 AF-predictors of *stage-1* are concatenated and used as input to train separate DNNs for each of the 5 AF-predictors. The AF-predictors developed in *stage-2* are called *MTL-1 AF-predictors*. Similarly, *MTL-2 AF-predictors* are developed in *stage-3* by concatenating the MFCCs and AFs from all the 5 AF-predictors of *stage-2* followed by training a separate DNN for each of the 5 AF-predictors.

table 8 shows the AF-EER of various AF-predictors developed using non-MTL (stage-1) and MTL (stage-2 and stage-3) approaches. The procedure for computation of AF-EER of AF-predictors is mentioned in section 5.3. AF-EERs of *non-MTL AF-predictors* are the same as the ones shown in table 5 and are shown here for comparison with the MTL-based AF-predictors. The MTL-1 and MTL-2 systems correspond to stage-2 and stage-3 blocks of figure 7, respectively.

The AF-EER of all the MTL-based AF-predictors has consistently reduced compared with the non-MTL AF-predictors. All the non-MTL AF-predictors have shown the highest AF-EERs while all the MTL-2 AF-predictors have shown the least AF-EERs. The AF-EERs of MTL-1 AF-predictors are between those of non-MTL and MTL-2 AF-predictors.

7.2 *Multi-PRSs using enhanced AFs from MTL-based AF-predictors*

We have combined the AFs predicted from the MTL-based AF-predictors to develop Multi-PRSs. From section 6, it can be observed that the AF-Tandem method is less complex and has better performance compared with LRA method of combining the AFs from various AF-predictors. Hence, we have used only AF-Tandem method for combining the AFs from various MTL-based AF-predictors. Table 9 shows the PERs of combined Multi-PRSs using both MTL- and non-MTL-based AF-predictors. We have considered only the results of All-AF-based and All-AF-PP-based Multi-PRSs using AF-Tandem method for combination. The results shown in second column indicate the non-MTL-based Multi-PRSs that are described in section 6.2. The values in the second column are taken from table 7 (AF-Tandem case) and are used for comparison with the results of MTL-based Multi-PRSs.

Multi-PRSs based on MTL-1 have not shown any improvement in their PERs; instead, both of them have a poorer performance compared with non-MTL Multi-PRSs by a very small margin of 0.1%. In case of MTL-2, *All-AF-based Multi-PRS* has shown a degraded performance of 0.1% PER whereas the *All-AF-PP-based* Multi-PRS has shown an improvement of 0.4% PER compared with non-MTL systems. The best performing MTL-2 Multi-PRS (31.9%) has a reduction of 0.4% PER compared with the best performing non-MTL Multi-PRS (32.3%). It is found that the *All-AF-PP-based* Multi-PRS using MTL-2 shows the least PER of 31.9% with an absolute reduction of 3.2% in the PER (9.1% reduction in relative PER) compared with baseline Multi-PRS.

8. Summary and conclusions

The baseline Multi-PRS is developed using 4 Indian languages – Kannada, Telugu, Bengali and Odia. MTL-based AF-predictors have better performance compared with non-MTL AF-predictors. The combination of AFs using AF-Tandem method performs better than that of LRA method. The *All-AF-PP-based* Multi-PRS based on MTL-2 AF-predictors has shown the least PER. The best performing predicted AFs have shown a reduction of 3.2% in absolute PER (9.1% reduction in relative PER), while the oracle AFs have shown an absolute reduction of 24.7% compared with baseline Multi-PRS. Given the remarkably low PER of $\sim 10\%$ for oracle-based Multi-PRS, it is concluded that there is much scope for enhanced prediction of AFs to improve the Multi-PRS to reach the performance of oracle AFs.

Acknowledgements

We thank Prof. B Yegnanarayana, Prof. K Sri Rama Murthy and Prof. R Kumaraswamy for providing Telugu and Kannada datasets.

References

[1] The International Phonetic Association 2007 *Handbook of the International Phonetic Association*. Cambridge University Press

[2] Stuker S, Metze F, Schultz T and Waibel A 2003 Integrating multilingual articulatory features into speech recognition. In: *Proceedings of INTERSPEECH*, pp. 1033–1036

[3] Manjunath K E and Sreenivasa Rao K 2017 Improvement of phone recognition accuracy using articulatory features. *Circuits, Systems, and Signal Processing* 37(2): 704–728

[4] Gerfen. 2011 *Phonetics theory* [online]. Available: http://www.unc.edu/g̃erfen/Ling 30Sp2002/phonetics.html, pages 251–257

[5] Narayanan S *et al* 2011 A multimodal real-time MRI articulatory corpus for speech research. In: *Proceedings of INTERSPEECH*, pp. 837–840

[6] Narayanan S *et al* 2014 Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America* 136(3): 1307–1311

[7] Lee S, Yildirim S, Kazemzadeh A and Narayanan S 2005 An articulatory study of emotional speech production. In: *Proceedings of INTERSPEECH*, pp. 497–500

[8] The Centre for Speech Technology Research, The University of Edinburgh. *MOCHA-TIMIT: MOCHA MultiCHannel Articulatory database: English* [online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

[9] Afshan A and Ghosh P K 2016 Better acoustic normalization in subject independent acoustic-to-articulatory inversion: benefit to recognition. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5395–5399

[10] Mitra V, Sivaraman G, Nam H, Espy-Wilson C and Saltzman E 2014 Articulatory features from Deep Neural Networks and their role in speech recognition. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3017–3021

[11] Kirchhoff K, Fink G A and Sagerer G 2002 Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication* 37: 303–319

[12] Frankel J, Magimai-Doss M, King S, Livescu K, Cetin O 2007 Articulatory feature classifiers trained on 2000 hours of telephone speech. In: *Proceedings of INTERSPEECH*

[13] Cetin O, Kantor A, King S, Bartels C, Magimai-Doss, Frankel J and Livescu K 2007 An articulatory feature-based tandem approach and factored observation modeling. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, p. IV-645

[14] Rajamanohar M and Fosler-Lussier E 2005 An evaluation of hierarchical articulatory feature detectors. In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 59–64

[15] Dusan S and Deng L 1998 Estimation of articulatory parameters from speech acoustics by Kalman filtering. In: *Proceedings of the CITO Researcher Retreat*, pp. 47–48

[16] Wakita H 1973 Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio, Speech, and Language Processing* 21(5): 417–427

[17] Dhananjaya N, Yegnanarayana B and Suryakanth V G 2011 Acoustic-phonetic information from excitation source for refining manner hypotheses of a phone recognizer. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[18] Corredor-Ardoy C, Lamel L, Adda-Decker M and Gauvain JL 1998 Multilingual phone recognition of spontaneous telephone speech. In: *Proceedings of ICASSP*, pp. 413–416

[19] Schultz T and Waibel A 2001 Language independent and language adaptive acoustic modeling for speech recognition. *Speech Communication* 35: 31–51

[20] Schultz T and Waibel A 1998 Multilingual and crosslingual speech recognition. In: *Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding*, pp. 259–262

[21] Schultz T and Kirchhoff K 2006 *Multilingual speech processing*. Academic Press

[22] Heigold G, Vanhoucke V, Senior A, Nguyen P, Ranzato M, Devin M and Dean J 2013 Multilingual acoustic models using distributed deep neural networks. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[23] Vu N T *et al* 2014 Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*

[24] Kumar C S, Mohandas V P and Haizhou L 2005 Multilingual speech recognition: a unified approach. In: *Proceedings of INTERSPEECH*, pp. 3357–3360

[25] Gangashetty S V, Sekhar C C and Yegnanarayana B 2005 Spotting multilingual consonant–vowel units of speech using neural network models. In: *Proceedings of the International Conference on Non-Linear Speech Processing (NOLISP)*, pp. 303–317

[26] Mohan A, Rose R, Ghalehjegh S H and Umesh S 2014 Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain. *Speech Communication* 56: 167–180

[27] Deng L 1997 Integrated-multilingual speech recognition using universal phonological features in a functional speech production model. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*

[28] Metze F 2005 *Articulatory features for conversational speech recognition*. PhD Thesis, Carnegie Mellon University

[29] Zhao Y, Zhao R, Wang X and Ji Q 2016 Multilingual articulatory features augmentation learning. In: *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2895–2899

[30] Livescu K *et al* 2007 Articulatory feature-based methods for acoustic and audio-visual speech recognition: summary from the 2006 JHU summer workshop. In: *Proceedings of the*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. IV-621–IV-624

[31] Black A W *et al* 2012 Articulatory features for expressive speech synthesis. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4005–4008

[32] Sahraeian R and Compernolle D V 2017 Crosslingual and multilingual speech recognition based on the speech manifold. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(12): 2301–2312

[33] King S, Frankel J, Livescu K, McDermott E, Richmond K and Wester M 2007 Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America* 121(2): 723–742

[34] Mermelstein P 1969 Computer simulation of articulatory activity in speech production. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 447–454

[35] Mermelstein P 1973 Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America* 53(4): 1070–1082

[36] Rubin P, Baer T and Mermelstein P 1981 An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America* 70(2): 321–328

[37] Mitra V, Wang W, Stolcke A, Nam H, Richey C, Yuan J and Liberman M 2013 Articulatory trajectories for large-vocabulary speech recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[38] Frankel J and King S 2007 Speech recognition using linear dynamic models. *IEEE Transactions on Audio, Speech, and Language Processing* 15(1): 246–256

[39] Zlokarnik I 1995 Adding articulatory features to acoustic features for automatic speech recognition. *The Journal of the Acoustical Society of America* 97(5): 3246–3246

[40] Mitra V *et al* 2014 Articulatory features from deep neural networks and their role in speech recognition. In: *Proceedings of ICASSP*, pp. 3017–3021

[41] Rasipuram R and Magimai.-Doss M 2016 Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Computer Speech and Language* 36: 233–259

[42] Stuker S, Schultz T, Metze F and Waibel A 2003 Multilingual articulatory features. In: *Proceedings of ICASSP*, vol. 1, pp. 144–147

[43] Schultz T 2002 GlobalPhone: a multilingual speech and text database developed at Karlsruhe university. In: *Proceedings of ICSLP*, Denver, CO, USA

[44] Ore B M 2007 *Multilingual articulatory features for speech recognition*. Master's Thesis, Wright State University

[45] Rasipuram R and Magimai-Doss M 2011 Improving articulatory feature and phoneme recognition using multitask learning. In: *Proceedings of Artificial Neural Networks and Machine Learning (ICANN)*, vol. 6791, pp. 299–306

[46] Muller M, Stuker S and Waibel A 2016 Towards improving low-resource speech recognition using articulatory and language features. In: *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pp. 1–7

[47] Muller M and Waibel A 2015 Using language adaptive deep neural networks for improved multilingual speech recognition. In: *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*

[48] Sahraeian R 2017 *Acoustic modeling of under-resourced languages*. PhD Thesis, Katholieke Universiteit Leuven (KU Leuven)

[49] Sahraeian R, Compernolle D V and de Wet F 2014 On using intrinsic spectral analysis for low-resource languages. In: *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*

[50] Dash D, Kim M, Teplansky K and Wang J 2018 Automatic speech recognition with articulatory information and a unified dictionary for Hindi, Marathi, Bengali, and Oriya. In: *Proceedings of INTERSPEECH*

[51] Manjunath K E, Rao K S, Jayagopi D B and Ramasubramanian V 2018 Indian languages ASR: a multilingual phone recognition framework with IPA based common phone-set, predicted articulatory features and feature fusion. In: *Proceedings of INTERSPEECH*

[52] *Development of prosodically guided phonetic engine for searching speech databases in Indian languages* [online]. http://speech.iiit.ac.in/svldownloads/pro_po_en_report/

[53] Kumar S B S, Rao K S and Pati D 2013 Phonetic and prosodically rich transcribed speech corpus in Indian languages: Bengali and Odia. In: *Proceedings of O-COCOSDA*, pp. 1–5

[54] Shridhara MV, Banahatti BK, Narthan L, Karjigi V and Kumaraswamy R 2013 Development of Kannada speech corpus for prosodically guided phonetic search engine. In: *Proceedings of O-COCOSDA*, pp. 1–6

[55] Madhavi M C, Sharma S and Patil H A 2014 Development of language resources for speech application in Gujarati and Marathi. In: *Proceedings of the IEEE International Conference on Asian Language Processing (IALP)*, vol. 1, pp. 115–118

[56] Sarma B D, Sarma M, Sarma M and Prasanna S R M 2013 Development of Assamese phonetic engine: some issues. In: *Proceedings of IEEE INDICON*, pp. 1–6

[57] Manjunath K E and Sreenivasa Rao K 2014 Automatic phonetic transcription for read, extempore and conversation speech for an Indian language: Bengali. In: *Proceedings of the IEEE National Conference on Communications (NCC)*

[58] Riedhammer K T, Bocklet T, Ghoshal A and Povey D 2012 Revisiting semi-continuous hidden Markov models. In: *Proceedings of ICASSP*, pp. 4721–4724

[59] Zhang X, Trmal J, Povey D and Khudanpur S 2014 Improving deep neural network acoustic models using generalized maxout networks. In: *Proceedings of ICASSP*, pp. 215–219

[60] Povey D *et al* 2011 The Kaldi Speech Recognition Toolkit. In: *Proceedings of the IEEE Workshop on ASRU*

[61] Manjunath K E, Jayagopi D B, Rao K S and Ramasubramanian V 2019 Development and analysis of multilingual phone recognition systems using Indian languages. *International Journal of Speech Technology*

[62] *Sclite Tool* [online]. http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

[63] Manjunath K E, Sreenivasa Rao K and Jayagopi D B 2017 Development of multilingual phone recognition system for Indian languages. In: *Proceedings of the IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*

[64] Erler K and Freeman G H 1996 An HMM-based speech recognizer using overlapping articulatory features. *Journal of Acoustic Society of America* 100(4): 2500–2513

[65] Ohman S E G 1965 Coarticulation in VCV utterances: spectrographic measurements. *Journal of Acoustic Society of America* 39(1): 151–168

[66] Ramachandran VR *Coarticulation knowledge for a text-to-speech system for an Indian language*. MS Thesis, Speech and Vision Laboratory, Indian Institute of Technology Madras, India

[67] Hermansky H, Ellis D P and Sharma S 2000 Tandem connectionist feature extraction for conventional HMM systems. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1635–1638

[68] Lal P and King S 2013 Cross-lingual automatic speech recognition using tandem features. *IEEE Transactions on Audio, Speech, and Language Processing* 21(12): 2506–2515

[69] Siniscalchi S M, Li J and Lee C 2006 A study on lattice rescoring with knowledge scores for automatic speech recognition. In: *Proceedings of INTERSPEECH*, pp. 517–520

[70] Rasipuram R and Magimai-Doss M 2011 Integrating articulatory features using Kullback–Leibler divergence based acoustic model for phoneme recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5192–5195

[71] Ketabdar H and Bourlard H 2008 Hierarchical integration of phonetic and lexical knowledge in phone posterior estimation. In: *Proceedings of ICASSP*, pp. 4065–4068

[72] Ketabdar H and Bourlard H 2010 Enhanced phone posteriors for improving speech recognition systems. *IEEE Transactions on Audio, Speech, and Language Processing* 18(6): 1094–1106

[73] Caruana R 1998 Multitask learning. In: *Learning to learn*. Boston, MA: Springer, pp. 95–133