Interspeech 2018
2-6 September 2018, Hyderabad

Indian languages ASR: A multilingual phone recognition framework with IPA based common phone-set, predicted articulatory features and feature fusion

Manjunath K E[1], *K. Sreenivasa Rao*[2], *Dinesh Babu Jayagopi*[1], *V Ramasubramanian*[1]

[1]International Institute of Information Technology - Bangalore (IIIT-B), Bangalore, India
[2]Indian Institute of Technology Kharagpur, Kharagpur, India

manjunath.ke@iiitb.org, ksrao@iitkgp.ac.in, {jdinesh,v.ramasubramanian}@iiitb.ac.in

Abstract

In this study, a multilingual phone recognition system for four Indian languages - Kannada, Telugu, Bengali, and Odia - is described. International phonetic alphabets are used to derive the transcription. Multilingual Phone Recognition System (MPRS) is developed using the state-of-the-art DNNs. The performance of MPRS is improved using the Articulatory Features (AFs). DNNs are used to predict the AFs for place, manner, roundness, frontness, and height AF groups. Further, the MPRS is also developed using oracle AFs and their performance is compared with that of predicted AFs. Oracle AFs are used to set the best performance realizable by AFs predicted from MFCC features by DNNs. In addition to the AFs, we have also explored the use of phone posteriors to further boost the performance of MPRS. 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 to baseline MPRS with MFCCs alone. The best performing system using predicted AFs has shown 2.8% reduction in absolute PER (8% reduction in relative PER) compared to baseline MPRS.

Index Terms: Indian languages ASR, multilingual framework, predicted AFs, feature fusion, deep learning

1. Introduction

There have been significant efforts in developing multilingual speech recognizers, a detailed description of which is given in [1] including issues, technologies and applications of multilingual speech recognition. Some of the notable work in this direction are that of [2], which develops a multilingual phone recognizer for spontaneous telephone speech for 4 languages - French, British English, German, and Castillan Spanish, [3] in which multilingual acoustic models are used to estimate the acoustic models for a new language in a fast and efficient way, and the design of a multilingual speech recognizer using GlobalPhone LVCSR dictation database is described [4].

A Multilingual Phone Recognition System (MPRS) is faced with the specific difficulty of having to arrive at the appropriate phone set based on which such a phonetic decoding can be done on input speech from any of the languages of interest. Such a common phone set has to have a coverage of all the phones occurring across the multiple languages. Our aim here is to develop a MPRS for 4 Indian languages and propose the use of International Phonetic Alphabet (IPA) based common multilingual phone-set which involves mapping acoustically similar phonetic units across languages to an underlying IPA unit. This is particularly appropriate considering that the IPA has strict one-to-one correspondence between symbols and sounds which makes it to be able to accommodate all the world's diverse languages.

While noting that no other multilingual effort has examined the use of IPA to derive a common phone-set labeling mechanism in the context of Indian languages, we also note that, multilingual speech recognition work using Indian languages has been limited to the following rather simplistic approaches - a syllable-based multilingual speech recognizer for 3 Indian languages Tamil, Telugu and Hindi [5], an isolated word recognition system for 2 linguistically similar Indian languages Hindi and Marathi [6] and, a bilingual phone recognizer for Tamil and Hindi [7]. Our work, based on the IPA transcription, represents an unifying framework generalizable to new languages easily.

The other main paradigmatic 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. The AFs can be continuous or discrete (see for example [8]), with the Mermelstein model [9, 10, 11] being a classic example of the continuous model. AFs have been consistently shown to improve speech recognition performance, such as in [12, 13, 14] (using continuous valued AFs) and in [15, 16, 17] (for the discrete valued AFs). AFs represent a higher degree of *invariance* and hence it is more appropriate to use them in multilingual tasks. With respect to use of AFs for multilingual speech recognition, the notable work are those of [18] where it was shown that the speech recognition performance can be improved by integrating the cross-lingual and multilingual AFs, [19] which showed that the inter-language variability can be compensated using AF detectors and [20] which used a multilayer perceptron (MLP) based estimation of multilingual AFs to improve the performance of multilingual systems. With regard to estimation of the AFs, [21] and [22] explored Deep Neural Networks (DNNs) as against the earlier work on use of MLP and [22] used DNN derived AFs for multilingual speech recognition. DNNs for multilingual speech recognition are reported in [23, 24, 25, 26].

In this paper, we focus on building a MPRS which can identify the phonetic units present in a given speech utterance independent of the language of the speech utterance. We also examine DNN based AF prediction 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. In essence, this work focuses on how best to arrive at a feature space (the AF parameter space) and the common phone set (the IPA set) that ensures enhanced *invariance* of the phonetic units amidst the increased variability due to the multilingual nature of the MPRS problem. This is perhaps the first of its effort in the context of Indian languages in several fronts, such as the use of IPA based description of the common phone set for MPRS, the use of DNN derived AFs as features with improved Phone Error Rate (PER) and establishing very low PERs (~10%) for *oracle*

10.21437/Interspeech.2018-2529

AFs, thereby setting the baseline performance achievable if AFs could be estimated accurately from speech directly or via other spectral representations.

The rest of the paper is organized as follows: Section 2 describes our experimental setup. Detailed description of development of MPRS, use of AFs, and the feature fusion is given in section 4. Section 5 provides the summary of the paper.

2. Experimental Setup

2.1. Speech Corpora

The speech corpora of 4 Indian languages namely, Kannada (KN), Telugu (TE), Bengali (BN) and Odia (OD) was collected as a part of consortium project titled *Prosodically guided phonetic engine for searching speech databases in Indian languages* supported by DIT, Govt. of India [27]. Speech corpora contains 16 bit, 16 KHz speech wave files along-with their IPA transcription [28]. The wave files contain read speech sentences of size between 3 to 10 seconds. Detailed description of the speech corpora is provided in [29, 30, 31, 32]. We have used a split of 80 : 20 for train and test data, respectively. 10% of training data is held out from the training and used as development set. Table 1 shows the statistics of the speech corpora.

Language	**# Speakers**		**Duration (in hours)**			
	M	**F**	**Train**	**Dev**	**Test**	**Total**
Kannada	7	9	2.80	0.33	0.76	3.89
Telugu	9	10	4.05	0.47	1.07	5.59
Bengali	20	30	3.42	0.40	0.99	4.81
Odia	14	16	3.58	0.36	0.97	4.91
Total	50	65	13.85	1.56	3.79	19.20

Table 1: *Statistics of Multilingual Speech Corpora.*

2.2. Training HMMs and DNNs

Initially flat-start initialization is used to build Context-Independent (CI) GMM-HMMs (referred as HMMs throughout). The alignments generated by the CI HMMs are used to initialise the training of Context-Dependent (CD) HMMs. This is further followed by training the CD DNN-HMMs (referred as DNNs throughout) using the alignments obtained from the CD HMMs. We have also trained CI DNNs using the alignments generated by the CI HMMs. The CI models are based on monophones, while the CD models are based on triphones. The mapping from phonetic context and the HMM-state index, to an emission probability density function is captured through acoustic-phonetic decision tree [33]. Number of Gaussians, number of transition states and number of transition ids depend on the number of phones and context being modelled. For example, baseline MPRS (in section 3) with 44 phones had 974 Gaussians, 132 transition states and 264 transition ids with CI HMMs, while the CD HMMs for the same system had 15039 Gaussians, 2078 transition states and 4156 transition ids.

DNNs with tanh non-linearity 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. Once all the hidden layers are added to the network, shrinking is performed after every 3 iterations, so as to separately scale the parameters of each layer. Mixing up was carried out in the halfway between the completion of addition of all the hidden layers and the end of training. Stability of the training is maintained through preconditioned affine components. Once the final iteration of training completes, the models from last 10 iterations are combined into a single model. Each input to DNNs uses a temporal context of 9 frames (4 frames on either side). The number of hidden layers for DNNs used in both AF-predictors and MPRSs are tuned, by adjusting the width of the hidden layers. It is found that the DNNs with 4 hidden layers are good for AF-predictors (in section 4.1), and DNNs with 5 hidden layers are suitable for MPRSs (in section 3). DNNs with dimensions of 432 input, 300 hidden, and 19860 output layer are used for training the baseline MPRS (in section 3). The total number of parameters of the DNNs range between 1.9 million to 2.0 millions based on the dimension of the features used. The size of the input layer depends on the dimension of features used for training the DNNs.

Bi-phone (phoneme bi-grams) language model is used for decoding. The language model weighting factor and acoustic scaling factor used for decoding the lattice are optimally determined using the development set to minimize the PER. DNNs training used in this study is similar to the one presented in [34]. All the experiments are conducted using the open-source speech recognition toolkit - Kaldi [35].

3. Development of Multilingual Phone Recognition Systems

A MPRS is developed using four Indian languages - KN, TE, BN, and OD. The common multilingual phone-set is derived by grouping the acoustically similar IPAs across languages together and selecting the phonetic units which have sufficient number of occurrences to train a separate model for each of them. The IPAs which do not have sufficient number of occurrences will be mapped to the closest linguistically similar phonetic units present in the common multilingual phone-set. The common multilingual phone-set thus derived contained 44 phones. We have also developed monolingual Phone Recognition Systems (PRSs) for KN, TE, BN, and OD languages using 36, 35, 34, and 36 phones, respectively. The 13-dimensional MFCCs [36] along-with their first and second order derivatives are computed using a frame-length of 25 ms with a frame-shift of 10 ms. Cepstral mean and variance normalization is applied per-speaker basis on MFCCs, followed by transformation using linear discriminant analysis. Both HMMs and DNNs are explored for developing PRSs under CD and CI settings.

Table 2 shows the PERs of monolingual and MPRSs. PER is computed by comparing the decoded phones with the reference phone labels. In Table 2, as we move from left to right PERs decrease in all the rows. This indicates that the CD models have lower PERs than CI models. In all the cases, DNNs have shown improved performance compared to HMMs. Since, the CD DNNs have shown least PERs in all the cases, we have used only CD DNNs in all our further experiments.

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

Table 2: *PERs of Monolingual and Baseline Multilingual Phone Recognition Systems developed using MFCCs.*

MPRS training uses the data shared from all the four languages. This makes the data used in the development of MPRS to have relatively higher number examples to be learnt by the DNNs and results in more accurate acoustic models compared to the monolingual systems, which are trained on a comparatively smaller amount of data. MPRS using CD DNNs outperform KN and BN monolingual PRSs. It is found that consonants are better modelled by KN and TE PRSs, while the vowels are more accurately modelled in BN and OD PRSs. MPRS takes the mutual advantage of all the languages and results in more accurate models for both consonants and vowels.

We have further analysed the causes for misclassifications at language and phone levels. If the number of examples contributed by a language towards training a phone are higher, then the misclassifications due to that specific phone from that particular language will be lower, and vice-versa. For example, the contribution of Odia language to the training data of */sh/* model of MPRS is only 0.5%. This indicates that the */sh/* phones occur very rarely in Odia. Although none of the */sh/* occurrences were present in the test data of Odia, but many fricatives (more prominently */s/*) are decoded as */sh/*. This increased the overall misclassification rate of */sh/*. Similarly, the contribution of Odia language to the training data of */ŋ/* is only 2.3%, which resulted in large misclassifications of Odia test utterances to */ŋ/*.

4. Articulatory Features for Multilingual Phone Recognition

We have used the DNNs for predicting the AFs from the speech signal. The predicted AFs and MFCCs are combined using two approaches namely - i) Lattice Rescoring Approach (LRA), ii) Combining AFs as Tandem features (AF-Tandem). Figure 1 shows the block diagram of combination of AFs using LRA. There are 3 stages in Figure 1. In the first stage, the AF predictors are developed to predict the AFs for five 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 MPRSs. Since, these MPRSs are developed using AFs and are arranged in tandem, we call them AF based tandem MPRSs. 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 MPRSs developed in the second-stage.

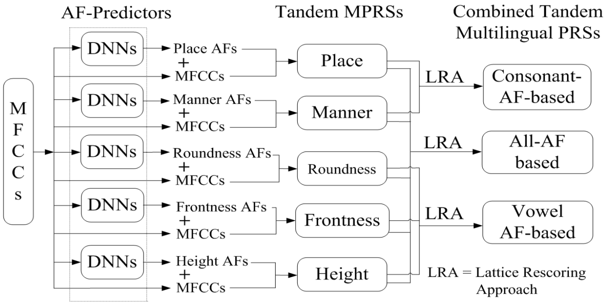

Figure 1: *Development of MPRS using Articulatory Features based on Lattice Rescoring Approach.*

In AF-tandem approach of combining the AFs, the estimated AFs from all the five AF groups are used as tandem features along-with MFCCs to develop MPRSs. Following subsections provide a detailed description.

4.1. Development of Articulatory Feature Predictors

AFs are predicted for five AF groups - place, manner, roundness, frontness, and height - using AF predictors. We have explored both DNNs and shallow neural networks having one hidden layer (FFNNs) to develop AF predictors. The frame-level AF labels required for training the models for each AF group are obtained by mapping the phone labels to AF label. MFCCs with 39 dimensions are used for training the AF-predictors. Table 3 shows the AF specification for different AF groups. AF-predictors are trained for classification of the features shown in Table 3. The cardinality indicates the number of classes in a AF group. The posterior probabilities generated by the AF-predictors represent AFs.

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

Table 3: *Articulatory Feature Specification for Different AF Groups (Crd = Cardinality).*

The performance of AF-predictors is evaluated by computing the frame-wise accuracy and the Mean Squared Error (MSE) between the predicted and oracle AFs. The frame-wise accuracy of each AF predictor is computed by comparing the predicted AF label of each frame with that of the actual AF label [16, 17]. Table 4 shows the framewise accuracy for various AF groups. The results of DNNs (5 hidden layers) and FFNNs (1 hidden layer) are shown separately. It is found that the performance of DNNs is much better compared to that of FFNNs for all the AF groups. We have also made similar observations on CI DNNs. This indicates that the use of DNNs is more beneficial compared to shallow neural networks for estimating the AFs. Hence, the DNN based AF-predictors are considered in all our experiments. *Roundness* AF group shows the highest accuracy, while the *height* AF group has shown least performance. We have also tried out *voicing* AF group with three classes - *voiced, unvoiced and silence*. But, the performance of *voicing* AF group was very poor (around 54%). This was due to the large number of misclassifications between silence and unvoiced.

AF Group	Framewise Accuracy (%)		Mean Squared Error (MSE)	
	DNN(5HL)	FFNN(1HL)	DNN(5HL)	FFNN(1HL)
Place	85.6	80.5	0.025	0.033
Manner	89.4	85.7	0.028	0.038
Roundness	90.8	87.1	0.037	0.055
Frontness	84.8	81	0.048	0.060
Height	80.5	77	0.051	0.059

Table 4: *Framewise Accuracy and Mean Squared Errors of AF-Predictors (HL = Hidden Layers).*

4.2. Development of AF based Tandem Multilingual Phone Recognition Systems

AFs predicted from each AF-predictor are augmented with MFCCs to develop a AF based tandem MPRS. This results in development of five AF based tandem MPRSs corresponding to five AF groups. 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 framelevel. The framelevel 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.

The PERs of AF based tandem MPRSs are shown in Table 5. The results are shown separately for predicted and oracle AFs. It is observed that the PERs of all the tandem MPRSs shown in Table 5 are superior than the baseline MPRS (see bold values in the MPRS row of Table 2). 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 MPRS has shown the highest reduction in PER, and *roundness* AF based tandem system has shown least reduction using predicted AFs. This is because *place* AF group has highest cardinality (i.e. 9), while the *roundness* has 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 to 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.

Features	**PER (%) of CD DNNs**	
	Predicted AFs	**Oracle AFs**
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

Table 5: *PERs of AF based Tandem Multilingual Phone Recognition Systems.*

4.3. 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, ii) AF-Tandem approach. In LRA approach, the lattices generated by the AF based tandem systems are combined using the lattice rescoring method [37]. The weighting factors required for LRA are tuned using development set. In AF-Tandem method of combination, the AFs are augmented as tandem features along-with MFCCs to develop MPRSs [16, 17]. The AFs derived from the consonant AF groups are combined to develop consonant-AF-based MPRS, while the vowel-AF-based MPRS is developed by combining the AFs from vowel AF groups. All-AF-based MPRS is developed by combining all the five AF-based tandem systems.

Table 7 shows the PERs of different AF-based MPRSs 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 to Vowel-AF-based, while the All-AF-based has higher PER reduction compared to Consonant-AF-based system. The PER of All-AF-based MPRS using oracle AFs is 22.3% lower than that of predicted AFs. Given the remarkably low PER of ~10% for oracle based MPRS, there is much scope for enhanced prediction of AFs to improve the MPRS to reach the performance of oracle AFs.

Further, we have also explored combining the Phone Posteriors (PPs) along-with all the predicted AFs to develop All-AF-PP-based MPRS. Similar to AFs, the PPs are predicted from the MFCCs using DNNs as described in [38]. All-AF-PP-based MPRS based on LRA has shown a PER of 32.6%, while the AF-Tandem method resulted in a PER of 32.3%. It is observed that the LRA method of combination has least PERs for consonant-AF-based, vowel-AF-based, and All-AF-based MPRSs, while the AF-Tandem method of combination has shown least PERs for All-AF-PP-based MPRS. 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 MPRS using oracle PPs.

Combined MPRSs	**Predicted AFs**		**Oracle AFs**	
	LRA	**AF-Tandem**	**LRA**	**AF-Tandem**
VAB	33.4	34.8	22.1	21.8
CAB	33.0	33.7	19.6	17.8
AAB	32.7	33.5	12.9	10.4

Table 6: *PERs of Combined Tandem Multilingual Phone Recognition Systems (VAB = Vowel-AF-based, CAB = Consonant-AF-based, AAB = All-AF-based).*

Combined MPRS	**Predicted AFs**	
	LRA	**AF-Tandem**
All-AF-PP-based MPRS	32.6	**32.3**

Table 7: *PERs of All-AF-PP-based Multilingual Phone Recognition Systems.*

The AF-Tandem method (through All-AF-PP-based MPRS) has shown the least PER of 32.3% with an absolute reduction of 2.8% in the PER (8% reduction in relative PER) compared to baseline MPRS. The PER of best performing MPRS (32.3%) is much better than the average of PERs of all monolingual PRSs (33.0%). The AF-Tandem method not only performs better than LRA but also has less complex structure than LRA. The time complexity of LRA is almost 5× higher than 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 MPRS has shown higher improvement in PERs compared to the vowel-AF-based MPRS. Since there are only 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 are higher than that of vowel AFs.

5. Summary and Conclusions

The baseline MPRS developed using CD DNNs has PER close to that of average of monolingual PRSs. The AFs predicted from DNNs are better than that of shallow neural networks. The combination of AFs using AF-Tandem method performs better than that of LRA method. The best performing predicted AFs have shown a reduction of 2.8% in absolute PER (8% reduction in relative PER), while the oracle AFs have shown an absolute reduction of 24.7% compared to baseline MPRS. Given the remarkably low PER of ~10% for oracle based MPRS, it is concluded that there is much scope for enhanced prediction of AFs to improve the MPRS to reach the performance of oracle AFs.

Acknowledgement : Thanks to Prof. B. Yegnanarayana, Prof. K. Sri Rama Murthy, and Prof. R. Kumaraswamy for providing Telugu and Kannada datasets.

6. References

[1] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Academic Press, 2006.

[2] C. Corredor-Ardoy et. al, "Multilingual phone recognition of spontaneous telephone speech," in *ICASSP*, 1998, pp. 413–416.

[3] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition," in *International Conference on Spoken Language Processing (ICSLP)*, 1998, pp. 1819–1822.

[4] ——, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998, pp. 259–262.

[5] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Spotting Multilingual Consonant-Vowel Units of Speech Using Neural Network Models," in *International Conference on Non-Linear Speech Processing (NOLISP)*, 2005, pp. 303–317.

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

[7] C. S. Kumar, V. P. Mohandas, and L. Haizhou, "Multilingual Speech Recognition: A Unified Approach," in *INTERSPEECH*, 2005, pp. 3357–3360.

[8] S. King, Frankel, Livescu, McDermott, Richmond and Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

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

[10] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53(4), pp. 1070–1082, 1973.

[11] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *The Journal of the Acoustical Society of America*, vol. 70, no. 2, pp. 321–328, 1981.

[12] V. Mitra, W. Wang, A. Stolcke, H. Nam, C, Richey, J. Yuan, and M. Liberman, "Articulatory Trajectories for Large-Vocabulary Speech Recognition," *ICASSP*, pp. 7145–7149, 2013.

[13] J. Frankel and S. King, "Speech recognition using linear dynamic models," *IEEE TASLP*, vol. 15, no. 1, pp. 246–256, 2007.

[14] I. Zlokarnik, "Adding articulatory features to acoustic features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3246–3246, 1995.

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

[16] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, O. Cetin, "Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech," in *INTERSPEECH*, 2007.

[17] O. Cetin, A. Kantor, S. King, C. Bartels, Magimai-Doss, J. Frankel, and K. Livescu, "An Articulatory Feature-Based Tandem Approach and Factored Observation Modeling," in *ICASSP*, vol. 4, 2007, pp. IV–645.

[18] S. Stuker, F. Metze, T. Schultz, and Alex Waibel, "Integrating Multilingual Articulatory Features Into Speech Recognition," in *INTERSPEECH*, 2003, pp. 1033–1036.

[19] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *ICASSP*, vol. 1, 2003, pp. 144–147.

[20] B. M. Ore, "Multilingual Articulatory Features for Speech Recognition," Master's thesis, Wright State University, 2007.

[21] V. Mitra et al., "Articulatory features from deep neural networks and their role in speech recognition," in *ICASSP*, 2014, pp. 3017–3021.

[22] M. Muller, S. Stuker, and A. Waibel, "Towards Improving Low-Resource Speech Recognition Using Articulatory and Language Features," in *International Workshop on Spoken Language Translation (IWSLT)*, 2016, pp. 1–7.

[23] M. Muller and A. Waibel, "Using language adaptive deep neural networks for improved multilingual speech recognition," *International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[24] G. Heigold et al., "Multilingual Acoustic Models Using Distributed Deep Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[25] Y. Miao and F. Metze, "Improving Low-Resource CD-DNN-HMM using Dropout and Multilingual DNN Training," *INTERSPEECH*, pp. 2237–2241, 2013.

[26] N. T. Vu et al., "Multilingual deep neural network based acoustic modeling for rapid language adaptation," *IEEE International Conference on*, 2014.

[27] Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages, *[online] : http://speech.iiit.ac.in/svldownloads/pro_po_en_report/.*

[28] The International Phonetic Association, *Handbook of the International Phonetic Association*. Cambridge University Press, 2007. [Online]. Available: https://www.internationalphoneticassociation.org/

[29] S. B. S. Kumar, K. S. Rao, and D. Pati, "Phonetic and Prosodically Rich Transcribed Speech Corpus in Indian languages : Bengali and Odia," in *O-COCOSDA*, 2013, pp. 1–5.

[30] M.V. Shridhara, B.K Banahatti, L. Narthan, V. Karjigi, and R. Kumaraswamy, "Development of Kannada speech corpus for prosodically guided phonetic search engine," in *O-COCOSDA*, 2013, pp. 1–6.

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

[32] B. D. Sarma, M. Sarma, M. Sarma, and S. R. M. Prasanna, "Development of Assamese Phonetic Engine: Some Issues," in *IEEE INDICON*, 2013, pp. 1–6.

[33] K. T. Riedhammer, T. Bocklet, A. Ghoshal, and D. Povey, "Revisiting semi-continuous hidden Markov models," in *ICASSP*, 2012, pp. 4721– 4724.

[34] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *ICASSP*, 2014, pp. 215–219.

[35] D. Povey et al., "The Kaldi Speech Recognition Toolkit," *IEEE Workshop on ASRU*, 2011. [Online]. Available: http://kaldi-asr.org/

[36] L. Rabiner, B. Juang, and B. Yegnanarayana, *Fundamentals of Speech Recognition*. Pearson Education, 2008.

[37] S. M. Siniscalchi, J. Li, and C. Lee, "A study on lattice rescoring with knowledge scores for automatic speech recognition," in *INTERSPEECH*, 2006, pp. 517–520.

[38] H. Ketabdar and H. Bourlard, "Hierarchical Integration of Phonetic and Lexical Knowledge in Phone Posterior Estimation," in *ICASSP*, 2008, pp. 4065–4068.