Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Speech Recognition with Articulatory Information and a Unified Dictionary for Hindi, Marathi, Bengali, and Oriya

Debadatta Dash[1], *Myungjong Kim*[1], *Kristin Teplansky*[1,2], *Jun Wang*[1,2]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
The University of Texas at Dallas, TX, USA

{debadatta.dash, myungjong.kim, kristin.teplansky, wangjun}@utdallas.edu

Abstract

Despite the continuous progress of Automatic Speech recognition (ASR) technologies, these systems for Indian languages are still in infancy stage due to a multitude of challenges involved, including resource deficiency. This paper addressed this challenge with four Indian languages, Hindi, Marathi, Bengali, and Oriya by integrating articulatory information into acoustic features, thereby compensating the low resource property of these languages. Articulatory movements were recorded during speech production using an electromagnetic articulograph and trained together with acoustic features to build automatic speech recognizers for these languages. Both speaker-dependent and -independent recognition experiments were conducted by adopting three ASR models: Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM), Deep Neural Network (DNN)-HMM, and Long Short Term Memory recurrent neural network (LSTM)-HMM. A cross-language similarity was discerned in both acoustic and articulatory domains in the pairs of Oriya-Bengali and Hindi-Marathi. Based on these observations, a multi-lingual, multi-modal speech recognizer was built by constructing a unified dictionary consisting of common and unique phonemes of all the four languages, which significantly reduced the phoneme error rates.

Index Terms: speech recognition, Procrustes matching, hidden Markov model, long short term memory networks

1. Introduction

Automatic Speech Recognition (ASR) enables a machine to recognize voice commands by emulating a voice pattern against acquired vocabulary. ASR has been recently commercially used for resource-rich languages such as English, Mandarin, and a few European languages. Research for resource-scarce languages such as Indian languages, however, is yet to gain momentum. India, the second largest populated country, has more than 5000 languages, of which 22 are official. Moreover, for the 70% of Indian population living in rural areas, having ASR enabled machine applications built in their native language would be another true progress of ASR.

Primary research on ASR for Indian languages has been focused on Hindi [1], the national and most common language of India. For Hindi, researchers have studied isolated word recognition [2], online speech to text engine [3], large vocabulary ASR [4], speeding up the statistical pattern classification [5], and connected digit recognition system [6]. For ASR in Bengali, phoneme recognition [7], Tr acoustic modeling [8] and SPHINX3 based Shruti-II [9] have been investigated. ASR for Marathi and Oriya is limited, including IVR [10] for Marathi and isolated digit recognition for Oriya [11].

Articulatory information has been proven effective in ASR as an additional source to acoustic features for English [12, 13]. Various technologies have been used to track articulatory motion such as ultrasound [14], surface electromyography [15], and electromagnetic articulograph (EMA) [16, 17]. Each technique is unique in terms of their advantages for collecting articulatory motion data. Comparatively, EMA is a more direct measure of flesh point artiuculatory movement, since it captures the 3-D motion of sensors adhered to the articulators (tongue and lips) [18] and hence has been used in this study. To our knowledge, adding articulatory information on top of acoustic features has rarely been studied for Indian languages.

Multi-lingual ASR has been recently studied to build an effective ASR system for resource-scarce languages. Most of these research are based on modeling common acoustic parameters across languages [19, 20, 21]. An alternative approach is to build a common phoneme set. For example, fast bootstrapping of large vocabulary continuous speech recognition systems with multilingual phoneme sets [22] and language adaptive acoustic modeling [23] for Europian languages, language dependent state clustering [24] for African languages and multi-lingual DNNs for global phones [25]. Although few have attempted to develop a multi-lingual ASR for Indian languages [1, 26], low data availability combined with phonological differences such as long and short vowels, lack of aspirated stops, aspirated consonants, and multi-occurrence of allophones makes its efficient establishment a daunting task [27]. To overcome these difficulties, a Unified Dictionary (UD) consisting both common and language-specific phonemes of the multiple languages may be useful.

In this paper, we built an automatic speech recognizer for Hindi, Marathi, Bengali and Oriya languages with articulatory information combined into acoustic features. We have used GMM-HMM and DNN-HMM based speech recognition models for Speaker-Dependent (SD) ASR experiment. Besides these two models, we also applied LSTM-HMM for Speaker-Independent (SI) ASR experiment. High phoneme level similarity was found in Hindi-Marathi language pair as well as in Oriya-Bengali. Hence, we further built a multi-lingual ASR with Unified Dictionary (UD) consisting phonemes of all the four languages.

2. Data Collection

2.1. Participants and Speech Task

Two native male speakers (mean age=27.5 years) participated in the data collection for all of the four languages. Neither any issues of speech, hearing, cognitive or language disorders from the participants were reported nor they had any family history of

10.21437/Interspeech.2018-2122

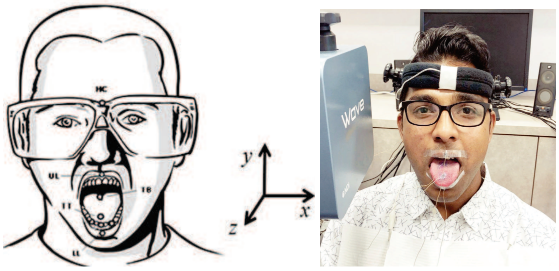

Figure 1: *Sensor labels and locations (described in the text).*

such disorders. Each subject had the speech task of saying 132 phrases (commonly used sentences for Alternative Augmented Communications (AAC), e.g., *I need to see a doctor.*) at their normal speaking rate in Hindi (*Subh dopahar*), Marathi (*Subh Dopar*), Bengali (*Subho Bikalo*) and Oriya (*Subha Aparanha*).

2.2. Articulatory Motion Tracking

The Wave system (Northern Digital Inc., Waterloo, Canada), a commercially available electromagnetic tongue and lip motion tracking device, was used to record the motion of the head, tongue, and lips. Four small sensors were attached to the surface of each articulator using dental glue (PeriAcryl 90, GluStitch) or tape, at the Tongue Tip (TT), Tongue Back (TB), Upper Lip (UL), and Lower Lip (LL) as in Figure 1. An additional sensor was attached to the middle of the forehead (Head Center, HC) for head motion correction. Our prior work has conveyed that this four-sensor set is an optimal set for recognition performance [28]. Hence, the flesh point three-dimensional articulatory movement data was chosen to be tracked and recorded from sensors placed at TT, TB, UL, and LL. The spatial precision of Wave for motion tracking is approximately 0.5 mm and the sampling rate of the recording was 100 Hz [18].

Prior to data analysis, translation and rotation of the HC sensor were subtracted from the motion data of the tongue and lip sensors to obtain head-independent articulatory data. Head translation and rotation removal was automatically done by the Wave system. Figure 1 illustrates the derived 3D Cartesian coordinates system, in which x is left-right direction; y is vertical and z is the front-back direction. Only y and z axes coordinates of the articulatory position with time were used for training as x-axis data, i.e, the lateral movement of articulators, is not very significant [18]. Figure 2 illustrates an example of tongue and lip motion for the word *"good"* in each of the four languages. The four languages have similar motion patterns for TB, but different for other sensors.

Acoustic data were collected synchronously with articulatory movement data by a built-in microphone in the Wave system with a sampling rate of 22050 Hz. 132 spoken utterances (average of $5,534$ phonemes/612 words) of each language were collected for analysis. The numbers of unique phonemes are 55, 66, 51 and 46 for Hindi, Marathi, Bengali, and Oriya, respectively. The average number of unique words in the four languages was 412.

2.3. Acoustic and Articulatory Features

Acoustic features are 39-dimensional MFCCs consisting of 13 static and their first and second derivatives with a frame size of 25 milliseconds and shift size of 10 milliseconds. We used 24 dimensional EMA data consisting of 8 static data (2 dimensions × 4 sensors) and their first and second order derivatives with shift size of 10 milliseconds. Mean normalization was done along each dimension as a default setting. The two feature vectors are concatenated and used as a final feature vector.

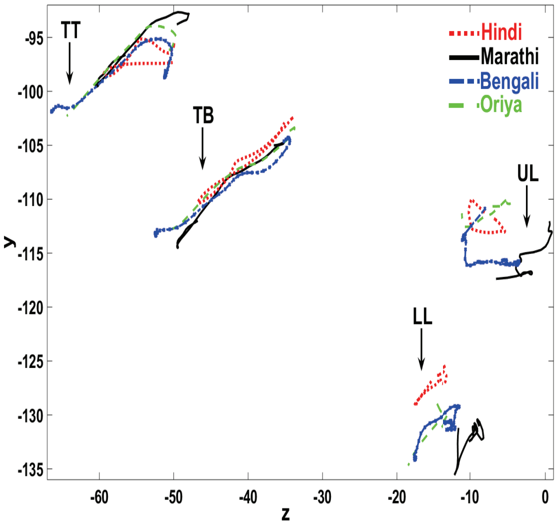

Figure 2: *Articulatory motion trajectories of four sensors (TT, TB, UL, and LL) for the word "good" in four languages.*

For SI approach, Procrustes matching based articulatory data normalization was performed to remove inter-talker physiological differences (tongue and lip orientation). In this method, two dimensional (i.e., y and z coordinates) movement data of TT, TB, UL, and LL were transformed into a normalized shape with a centroid at the origin $(0, 0)$. The centroids of the UL and LL formed a vertical line [29].

3. Methods

ASR performance using acoustic features with and without articulatory information were compared in this study. Three unique and different approaches such as GMM-HMM, DNN-HMM and LSTM-HMM were used for ASR analysis. For SD-ASR, due to low data availability, only the first two approaches were considered whereas for SI all three of them were used for analysis. Detailed configurations of the features and the ASR models are summarized in Table 1.

3.1. GMM-HMM

We trained with monophone GMM-HMM with different number of states (170 for Hindi, 203 for Marathi, 158 for Bengali and 143 for Oriya) for the four languages based on their unique number of existing phonemes. 3 states for each phone and 5 states for silence were taken. A total of 8 Gaussians per state were considered and a 3 state left to right HMM was used with a training method of Maximum Likelihood Estimation (MLE).

3.2. DNN-HMM

DNN-HMM is typically trained with multiple frames of speech features to produce posterior probabilities over HMM states as output for decoding. We trained the DNN with 5 hidden layers (optimal number based on experimentation) with each layer

Table 1: *Experimental setup*

Components	Details
Acoustic feature	
Feature vector	MFCC+ Δ + $\Delta\Delta$ Dimension = 39 (3 $\times$ 13)
Sampling rate	22050 Hz
Window length	25ms
Articulatory feature	TT, TB, UL, LL
Feature vector	8 sensors + Δ + $\Delta\Delta$ Dimension = 24 (3 $\times$ 8)
Concatenated feature	
Feature vector	MFCC + Δ + $\Delta\Delta$ + 8 sensors + Δ + $\Delta\Delta$ Dimension = 63(39 + 24)
Common	
Frame rate	10 ms
LSTM-HMM	
Input layer dimension	63 (24+39)
Output layer dimension	monophone 170 for Hindi 203 for Marathi 157 for Bengali 143 for Oriya 290 for Unified Lexicon
Number of LSTM cell units	320 per hidden layer
Depth	2 forward hidden layers
Training method	BPTT
Language model	Bigram phoneme LM
Metric of ASR	Phoneme error rates
Data sampling for training	6 fold cross-validation

consisting 128 nodes. The input layer was designed to take 9 frames at a time (4 previous+1 current+4 succeeding) with feature vector dimension of 216 (9$\times$24) when trained only using articulatory features, 351 (9$\times$39) for acoustic features and 567 (9$\times$(39+24)) when both were concatenated. The output layer of DNN was fixed and based on the language (170 for Hindi, 203 for Marathi, 158 for Bengali and 143 for Oriya). The parameters were initialized using layer-by-layer pre-training based on restricted Boltzmann machines (RBMs) and the network was trained using backpropagation [30].

3.3. LSTM-HMM

Long Short Term Memory (LSTM) recurrent neural networks contain memory blocks with a set of recurrently connected subnets [31], built similar to recurrent neural networks (RNN) by replacing the non-linear units with memory blocks in the hidden layers. These memory blocks help LSTM to overcome the vanishing gradient problem of RNN. Hence, LSTM models have been widely used in the area of ASR [29, 32]. In our model, each layer contained 320 cell units and parameters were trained using Back Propagation Through Time (BPTT).

The bigram phoneme language model was used for the phoneme sequence recognition. The training and decoding were performed using the Kaldi speech recognition toolkit [33]. Phoneme Error Rates (PERs) were used as the performance measure of Indian speech recognition. Six-fold cross-validation was used to perform SI phoneme recognition in the experiment.

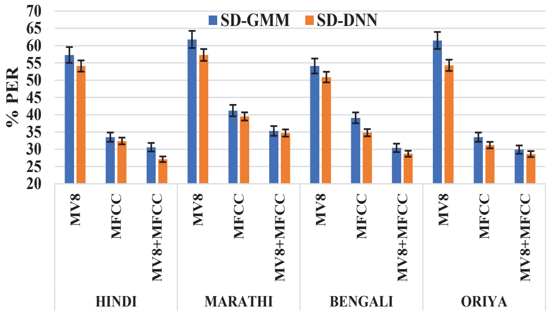

Figure 3: *Performance improvement by adding articulatory information in speaker-dependent ASR for Indian languages.*

3.4. Multi-Lingual ASR: UD-LSTM-HMM

Multilingual phoneme set was prepared from monolingual models by combining acoustically similar phones to form a Unified Dictionary of the four languages. It was assumed that these acoustically similar phones are also similar in their articulatory representations across Indian languages. Beside few Perso-Arabic scripts (Kasmiri, Urdu and Sindhi), all other scripts for Indian languages have been originated from the ancient Brahmi script. Hence, they are expected to have some common phonetics similarity. From Indic TTS-CLS [34] it is evident that there are 8 vowels and 33 consonants that are common in the four languages. The Unified Dictionary was built by defining a new dictionary that consists all unique phonemes in the four languages - with a total of 95 phonemes having 41 common phonemes and 54 unique (language-specific) phonemes. We trained SI-LSTM-HMM model on this unified dictionary with an output layer dimension of 290 (95 phonemes $\times$ 3 states + 5 states (silence)) to find the monphone based PERs.

4. Results and Discussion

4.1. Speaker-dependent recognition

Figure 3 shows the phoneme recognition performance of acoustic features only (MFCC), articulatory features only (MV8), and their combination (MV8+MFCC) on the speaker-dependent GMM- and DNN-based ASR systems for four Indian languages. Adding articulatory information on top of acoustic features significantly improved the ASR performance (PER reduction) for all the four Indian languages. This observation indicated that lip and tongue movement data contain complementary information to MFCC in phoneme recognition. We also investigated the performance using articulatory information only. As shown in Figure 3, when only articulatory data was used, the performance was consistently poorer than using MFCC. This finding is expected because articulatory features alone contain less information than speech acoustics (e.g., lack of voice).

In comparison, DNNs slightly outperformed GMM for all four languages (See Figure 3), although DNN typically requires a larger training data set for effective performance. The average Phoneme Error Rate reduction was about 3% than GMM-HMM approach. We obtained an average PER of 26.2%, 34.1%, 28.4% and 27.7% for Hindi, Marathi, Bengali and Oriya languages, respectively, using DNN-HMM based ASR recognition even when the training data was approximately 20 minutes for each language. The high PER in Marathi language (34.1%) is probably due to the presence of larger number of unique

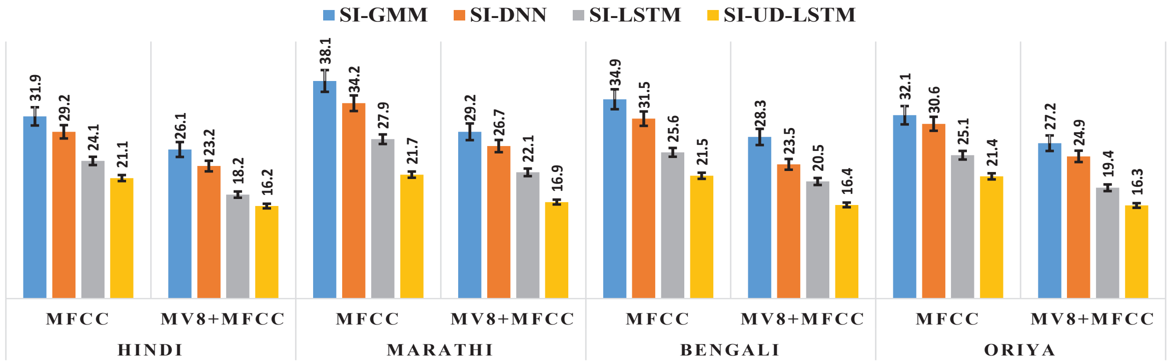

Figure 4: *Performances of different speaker-independent ASR models for Indian languages using acoustic and articulatory features.*

Table 2: *Performances (in %PER) for cross-lingual ASR*

Testing	Training				
	Hindi	Marathi	Bengali	Oriya	UD
Hindi	**18.2**	42.3	62.5	56.2	**16.1**
Marathi	49.8	**22.1**	69.2	67.3	**16.9**
Bengali	58.6	66.4	**20.5**	47.3	**16.4**
Oriya	59.3	63.4	46.6	**19.4**	**16.2**

phonemes in this language than the rest.

4.2. Speaker-independent recognition

Figure 4 shows the performance of SI GMM-, DNN-, LSTM and UD-LSTM-based speech recognizers for the four Indian languages, respectively, using MFCCs only as well as combined with articulatory data. Due to the effectiveness of Procrustes matching based articulatory normalization across speakers, both GMM and DNN based recognition approaches performed better than speaker-dependent case. As evident from Figure 4, an average of 5% PER improvement was discerned due to Procrustes matching itself. Similarly, for SI study, the inclusion of articulatory features enhanced the ASR performance by 6% PER reduction. LSTM model due to its inherent characteristics of sequential learning outperformed both GMM and DNN-based ASR approaches. This observation confirmed that LSTM neural network can be used for small sized data and maybe a better speech recognition architecture for Indian languages than DNNs.

Limitation. Although the results are promising, our data size is small and no female speakers were included. It is still unclear if the findings can be generalized to a larger number of Indian speakers include both males and females.

4.3. Recognition with a Unified Dictionary

To investigate the cross-language similarity, we performed a criss-cross training-testing procedure among the four languages in a speaker-independent way. We trained our model with the dictionary built on phonemes of Hindi language and tested with the other three language data. The same procedure was repeated for Bengali, Marathi and Oriya. Recognition performance of this experimentation is displayed in Table 2. The %PERs for the pairs of Hindi-Marathi and Bengali-Oriya were found to be comparatively less, which indicated the phoneme level similarity in-between these pairs. The same pattern was also observed in articulatory space as shown in Figure 2. Motion contours of tongue and lip movement for the same stimuli were found to be visually similar for Hindi and Marathi as well as for Bengali and Oriya. These results motivated for the construction of a multi-lingual ASR built on a Unified Dictionary.

The rightmost column in Table 2 indicates that %PER was significantly reduced with a Unified Dictionary, giving PERs about 16% for these languages. A possible explanation is due to the integration of all the phonemes in one dictionary, the training data set increased up to four times than the prior approaches. This resulted in repeated training of similar phonemes along with the corresponding articulatory movements for similar words. An example would be, the word *"good"* has the phoneme representation when translated into Hindi: $s - u - b^h$, in Marathi: $s - u - b^h$, in Bengali: $s - u - b^h - o$ and in Oriya: $s - u - b^h - ax$. Hence, with the Unified Dictionary, the phonemes s, u, b^h were trained four times for a single phrase and hence resulted in a reduced PER.

5. Conclusions

The results suggest that adding articulatory information on top of acoustics may improve the ASR performance for Indian languages. It also encourages for a multi-lingual approach to be more suitable for Indian languages due to the inter-language phonetic similarity. The experimental results also confirm on the existence of a high level of acoustic and articulatory similarity between Hindi and Marathi as well as between Bengali and Oriya. Moreover, LSTM neural network outperformed the GMM and standard DNN in our experiments. Although this study only included four languages, the results propose that a multilingual approach with added articulatory data can be generalized for other Indian languages. Due to the logistic difficulty of tongue motion data collection, (quasi-) articulatory data can be obtained from acoustic data based on a speaker-independent inverse mapping (e.g., [35]) in practice.

6. Acknowledgments

This work was supported by the National Institutes of Health (NIH) under award number R03 DC013990 and by the American Speech-Language-Hearing Foundation via a New Century Scholar Research Grant. We thank Beiming Cao, Dinesh Dash and the volunteering participants for helping in data collection. We also acknowledge Dr. Ted Mau for his valuable support.

7. References

[1] P. Lavanya, P. Kishore, and G. T. Madhavi, “A simple approach for building transliteration editors for Indian languages,” *Journal of Zhejiang University-SCIENCE A*, vol. 6, no. 11, pp. 1354–1361, Nov 2005.

[2] U. G. Patil, S. D. Shirbahadurkar, and A. N. Paithane, “Automatic Speech Recognition of isolated words in Hindi language using MFCC,” in *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, Dec 2016, pp. 433–438.

[3] B. Venkataramani, “SOPC-based speech-to-text conversion,” in *Nios II Embedded Processor Design Contest - Outstanding Designs*, 2006.

[4] M. Kumar, N. Rajput, and A. Verma, “A large-vocabulary continuous speech recognition system for Hindi,” *IBM Journal of Research and Development*, vol. 48, no. 5.6, pp. 703–715, Sep 2004.

[5] R. K. Aggarwal and M. Dave, “Using gaussian mixtures for hindi speech recognition system,” in *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2011.

[6] A. Mishra, M. Chandra, A. Biswas, and S. N Sharan, “Robust features for connected Hindi digits recognition,” *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, no. 2, Jun 2011.

[7] M. Kotwal, M. Shahadat Hossain, F. Hassan, G. Muhammad, M. Nurul Huda, and C. Mofizur Rahman, “Bangla phoneme recognition using hybrid features,” *Conference: Electrical and Computer Engineering (ICECE)*, pp. 718 – 721, Jan 2011.

[8] P. Banerjee, G. Garg, P. Mitra, and A. Basu, “Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali,” in *2008 19th International Conference on Pattern Recognition*, Dec 2008, pp. 1–4.

[9] S. Mandal, B. Das, and P. Mitra, “Shruti-ii: A vernacular speech recognition system in Bengali and an application for visually impaired community,” in *2010 IEEE Students Technology Symposium (TechSym)*, Apr 2010, pp. 229–233.

[10] S. Gaikwad and D. Gawali, “Polly clinic inquiry system using IVR in Marathi language,” *International Journal of Machine Intelligence*, vol. 3, Nov 2011.

[11] S. Mohanty and B. Kumar Swain, “Markov model based Oriya isolated speech recognizer-an emerging solution for visually impaired students in school and public examination,” Mar 2018.

[12] A. A. Wrench and K. Richmond, “Continuous speech recognition using articulatory data,” in *Sixth International Conference on Spoken Language Processing (ICSLP)*, vol. 4, 2000, pp. 145–148.

[13] K. Kirchhoff, G. A. Fink, and G. Sagerer, “Combining acoustic and articulatory feature information for robust speech recognition,” *Speech Communication*, vol. 37, no. 3, pp. 303 – 319, 2002.

[14] B. Denby, J. Cai, P. Roussel, G. Dreyfus, L. Crevier-Buchman, C. Pillot-Loiseau, T. Hueber, and G. Chollet, “Tests of an interactive, phrasebook-style, post-laryngectomy voice-replacement system,” in *17th International Congress of Phonetic Sciences (ICPhS 2011)*, Hong Kong, China, Aug. 2011, p. 1.

[15] Y. Deng, J. Heaton, and G. Meltzner, “Towards a practical silent speech recognition system,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Jan 2014, pp. 1164–1168.

[16] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, “Development of a (silent) speech recognition system for patients following laryngectomy,” *Medical Engineering & Physics*, vol. 30, no. 4, pp. 419 – 425, 2008.

[17] J. Wang, A. Samal, and J. Green, “Preliminary test of a real-time, interactive Silent Speech Interface based on Electromagnetic Articulograph,” *ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies*, Jun 2014.

[18] J. Wang, J. R. Green, A. Samal, and Y. Yunusova, “Articulatory distinctiveness of vowels and consonants: A data-driven approach,” *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[19] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, “Multilingual acoustic models using distributed Deep Neural Networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8619–8623.

[20] A. Ghoshal, P. Swietojanski, and S. Renals, “Multilingual training of Deep Neural Networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7319–7323.

[21] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual Deep Neural Network with shared hidden layers,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7304–7308.

[22] T. Schultz and A. Waibel, “Fast bootstrapping of LVCSR systems with multilingual phoneme sets,” in *EUROSPEECH*, 1997.

[23] T. S. et al, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Communication*, vol. 35, no. 1, pp. 31 – 51, 2001.

[24] T. Niesler, “Language-dependent state clustering for multilingual acoustic modelling,” *Speech Communication*, vol. 49, no. 6, pp. 453 – 463, 2007.

[25] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, “Multilingual Deep Neural Network based acoustic modeling for rapid language adaptation,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7639–7643.

[26] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, “Speech recognition and keyword spotting for low resource languages: Babel project research at cued,” in *Proc. Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*, 2014, pp. 16–23.

[27] A. H. Unnibhavi and D. S. Jangamshetti, “A survey of speech recognition on south Indian languages,” in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, Oct 2016, pp. 1122–1126.

[28] J. Wang, A. Samal, P. Rong, and J. R. Green, “An optimal set of flesh points on tongue and lips for speech-movement classification,” *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 15–26, 2016.

[29] M. Kim, B. Cao, T. Mau, and J. Wang, “Speaker-Independent silent speech recognition from flesh-point articulatory movements using an lstm neural network,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, 2017.

[30] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep Neural Networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[31] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–80, Dec 1997.

[32] H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory recurrent neural network architectures for large scale acoustic modeling,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and K. Vesely, “The Kaldi speech recognition toolkit,” in *Proc. IEEE ASRU*, Waikoloa, USA, 2011, pp. 1–4.

[34] R. Boothalingam, L. Christina, A. Gladston, S. Solomi V, M. Kumar Nandwana, A. Prakash, A. Shanmugam, R. K. Kalyanaraman, K. Prahallad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. Murthy, “A common attribute based unified HTS framework for speech synthesis in Indian languages,” Aug 2013.

[35] S. Hahm and J. Wang, “Parkinson’s condition estimation using speech acoustic and inversely mapped articulatory data,” in *Proc. of INTERSPEECH*, 2015, pp. 513–517.