Study of Multilingual Phone Recognition using Indian Languages

Manjunath K E (Roll No : PH2014014)

International Institute of Information Technology - Bangalore, India

manjunath.ke@iiitb.org

Abstract

In this work, multilingual phone recognition using six Indian languages - Kannada, Telugu, Bengali, Odia, Urdu, and Assamese - is studied. International phonetic alphabets (IPA) based transcription is used to develop Multilingual Phone Recognition System (MPRS). MPRS is developed using the state-of-the-art DNNs. The language independent nature of articulatory features (AFs) is exploited to improve the PER of MPRS. The use of predicted AFs has resulted 2.8% reduction in absolute PER (8% reduction in relative PER) compared to the baseline MPRS. The advantages of MPRS in case of code-mixing scenario is studied using the code-mixed sentences of Kannada and Urdu languages. It is shown that the proposed MPRS performs better than the traditional approach for multilingual phone recognition that uses a two-stage approach consisting of a language identification block followed by a phone recognition block. It is also shown that the proposed MPRS is more advantageous in case of code-mixed phone recognition.

1. Key Ideas of the Thesis

- To develop a Multilingual Phone Recognition System (MPRS) using the International Phonetic Alphabets (IPA) based transcription of four Indian languages - Kannada, Telugu, Bengali, and Odia.
- To explore Articulatory Features (AFs) to improve the phone error rate of MPRS.
- To explore deep neural networks to derive articulatory features and to develop MPRS.
- To demonstrate the advantages of MPRS by comparing the performance MPRS with that of traditional two-stage phone recognisers (having a language identification block and a phone recognition block) for multilingual phone recognition.
- To demonstrate that MPRS has advantages in case of code-mixed phone recognition.

2. Motivation

India has 22 constitutionally recognised major languages, each of which is spoken by more than a million native speakers. Given the large number of languages, a study related to the development of MPRS is very essential in the context of Indian languages. Based on the origins of the languages, Indian languages are further sub-divided into four majour language families namely : Indo-Aryan, Dravidian. Austric, and Tibeto-Burman. Since the languages within a language group are closely related and linguistically similar, it will be more relevant to explore multilingual phone recognition for each language family and analyze their behaviour.

Given the large number of Indian languages, the studies exploring the multilingual phone recognition using Indian Languages are very limited. None of the multilingual efforts have examined the use of IPA [1] to derive a common phone-set labelling mechanism in the context of Indian languages. The existing studies related to multilingual speech recognition work using Indian languages have been limited to the following simplistic approaches - a syllable-based multilingual speech recognizer for 3 Indian languages Tamil, Telugu and Hindi [2], an isolated word recognition system for 2 linguistically similar Indian languages Hindi and Marathi [3] and, a bilingual phone recognizer for Tamil and Hindi [4]. Motivated by this, we would like to study the multilingual phone recognition in the context of Indian languages using International Phonetic Alphabets (IPA). The objective of the current study is to develop a unifying framework for development of MPRS using IPAs that could be generalizable to any of new languages.

The traditional approach for multilingual phone recognition consists of two stages - Language Identification (LID) block in the first-stage followed by a Monolingual phone recognisers in the second-stage. We compare the performance of traditional MPRSs with the proposed MPRSs and show that the proposed approach outperforms the traditional method. We address the problem of efficient phone recognition for code-switched speech for 2 Indian languages Kannada and Urdu. Few works related to code-mixing are reported in [5, 6, 7, 8]. The proposed multilingual phone recognition applied to code-mixing scenario is superior to a more conventional front-end LID-switched monolingual phone recognition trained individually on each of the languages involved in the code-switched speech.

3. Major contributions

The MPRS is developed using four Indian languages - Kannada (KN), Telugu (TE), Bengali (BN) and Odia (OD). The speech corpora described in [9, 10, 11, 12] is used in this study. 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 features extracted from the input signal along with their phonetic labels are used for training the HMMs/DNNs. A multilingual decoder is used for decoding the phones present in the test utterance. Some of the notable works in this direction are reported in [13, 14, 15, 16]. The use of DNNs for multilingual speech recognition are reported in [17, 18, 19, 20].

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

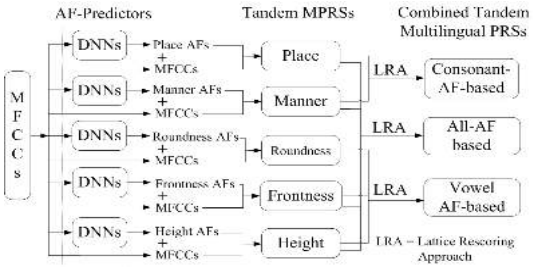

Figure 1: *Development of MPRS using Articulatory Features based on Lattice Rescoring Approach.*

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. 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. The AFs for multilingual speech recognition are rported in [21, 22, 23, 24, 25]

The traditional approach for multilingual phone recognition has two blocks - i) Language identification (LID), ii) phone recognition by monolingual PRS based on the language identified. We have explored SVMs [26] [27] to develop the LID block. This part of the study uses two other languages - Assamese (AS) and Urdu (UR) taken from the same speech corpora - in addition to the four languages described above.

4. Discussion of Results

Table 1 shows the PERs of monolingual and baseline MPRSs. The PERs of monolingual PRSs are better than baseline MPRS in case of HMMs. The reductions in PERs of monolingual PRSs using DNNs are not as high as that of baseline MPRS, which has a reduction of 14.3% absolute error in PER compared to HMMs. This indicates that MPRS using the data shared by four languages has 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 relatively smaller amount of data. MPRS using DNNs outperform KN, TE, and BN monolingual PRSs. This clearly shows the advantage of using DNNs for developing MPRSs. We 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.

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 1: *PERs of Monolingual and Multilingual Phone Recognition Systems developed using MFCCs.*

Table 2 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) [28] along-with all the predicted AFs to develop All-AF-PP-based MPRS. 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%. 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.

Combined MPRSs	Predicted AFs		Oracle AFs	
	LRA	AF-Tan	LRA	AF-Tan
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

Table 2: *PERs of MPRSs using Articulatory Features.*

Table 3 shows the LID accuracy (in %) for various languages using SVMs. SVMs LID classifiers are developed using MFCCs and i-vectors. It is found that SVM classification using i-vectors outperform MFCCs. Hence, we have considered only the i-vector based LID results in all our future experiments. LID accuracy decreases as no. of languages increase.

Languages	Features used for LID	
	MFCCs	i-vectors
KN-BN-OD-UR	91.16	97.98
KN-TE-BN-OD-UR	74.76	96.22
KN-TE-BN-OD-UR-AS	71.19	96.00

Table 3: *LID Accuracy (%) using SVMs.*

Table 4 shows the comparison of the performance of proposed MPRS with that of the traditional MPRSs. It is found that the proposed MPRS outperforms the traditional approach for multilingual phone recognition. This is more clear and evident with the increased number of languages. The proposed MPRS has the performance that is very close to that of Oracle MPRS. The use of MPRS has an additional advantage of decoding more number of phones than the number of phones decoded by the traditional approach.

Languages	MPRS Approach		Oracle
	Proposed	Traditional	
KN,BN,OD,UR	34.00	35.5	33.73
KN,TE,BN,OD,UR	33.40	35.6	33.12
KN,TE,BN,OD,UR,AS	35.20	37.9	35.30

Table 4: *Comparison: PERs of Proposed and Traditional MPRS*

Table 5 shows the comparison of proposed and traditional MPRSs for phone recognition in case of code-mising scenario. The traditional approach suffers due to a trade-off between two conflicting factors the need for short windows for detecting code-switching at a high time resolution and the need for long windows needed for reliable LID which limits the overall performance of the traditional system that suffers with high PERs at small windows (poor LID performance) and mismatched decoding conditions at long windows (due to poor code-switching detection time resolution). However, the proposed MPRS, by virtue of not having to do a front-end LID switching and by using a multi-lingual phone-set, is not constrained by these conflicting factors and hence performs effectively on code-switched speech, offering low PERs than the traditional system.

Languages	Proposed MPRS	Traditional MPRS		
		1-sec	2-sec	3-sec
KN,UR	32.7	42.64	38.47	36.58
KN,BN,OD,UR	31.5	43.93	38.81	36.67
KN,TE,BN,OD,UR	32.5	45.92	39.42	37.46
KN,TE,BN,OD,UR,AS	31.9	45.92	39.53	37.99

Table 5: *Comparison in Code-mixed Scenario.*

Acknowledgement : I thank my PhD supervisors Prof. V Ramasubramanian and Prof. Dinesh Babu Jayagopi.

5. References

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

[2] 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.

[3] 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.

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

[5] S. Ford. Language mixing among bilingual children. [Online]. Available: http://www2.hawaii.edu/ sford/research/mixing.htm

[6] J. F. Kroll and A. M. B. De Groot, Ed., *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press, 2005.

[7] R. R. Heredia and J. Altarriba, "Bilingual language mixing: Why do bilinguals code-switch?" *Current Directions in Psychological Science*, vol. 10, pp. 164–168, 2001.

[8] L. Jorschick, A. E. Quick, D. Glasser, E. Lieven, and M. Tomasello, "GermanEnglish-speaking childrens mixed NPs with correct agreement," *Bilingualism: Language and Cognition*, vol. 14, no. 2, pp. 173–183, 2011.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

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

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

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

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

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

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

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

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

[25] 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.

[26] M. Li, H. Suo, X. Wu, P. Lu, and Y. Yan, "Spoken Language Identification Using Score Vector Modeling and Support Vector Machine," in *Interspeech*, 2007, pp. 350–353.

[27] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language recognition with support vector machines," *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pp. 41–44, 2004.

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