Multilingual Phone Recognition: Comparison of Traditional versus Common Multilingual Phone-Set Approaches and Applications in Code-Switching

K. E. Manjunath[1,3(✉)], K. M. Srinivasa Raghavan[1], K. Sreenivasa Rao[2], Dinesh Babu Jayagopi[1], and V. Ramasubramanian[1]

[1] International Institute of Information Technology - Bangalore (IIIT-B), Bangalore, India
{manjunath.ke,srinivasaraghavan.km}@iiitb.org,
{jdinesh,v.ramasubramanian}@iiitb.ac.in
[2] Indian Institute of Technology Kharagpur, Kharagpur, India
ksrao@iitkgp.ac.in
[3] U. R. Rao Satellite Centre, ISRO, Bangalore, India

Abstract. We propose a multilingual phone recognition system using common multilingual phone-set (Multi-PRS) derived from IPA based labelling convention, which offers seamless decoding of the code-switched speech. We show that this approach is superior to a more conventional front-end language-identification (LID)-switched monolingual phone recognition (LID-Mono) trained individually on each of the languages present in multilingual dataset. The state-of-the-art i-vectors are used to perform LID. We address the problem of efficient speech recognition for bilingual code-switching. We analyse the differences between LID-Mono and proposed Multi-PRS, by showing that the LID-Mono 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 language identification - which limits the overall performance of the LID-Mono 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). We show that the Multi-PRS, by virtue of not having to do a front-end LID switching and by using a multilingual phone-set, is not constrained by these conflicting factors and hence performs effectively on code-switched speech, offering low PERs than the LID-Mono system.

Keywords: Code-switched ASR · Indian language ASR · Multilingual ASR · LID-switched monolingual ASR

S. M. Thampi et al. (Eds.): SIRS 2019, CCIS 1209, pp. 75–86, 2020.
https://doi.org/10.1007/978-981-15-4828-4_7

1 Introduction

The traditional approach for multilingual phone recognition uses front-end language-identification (LID)-switched monolingual phone recognition (LID-Mono) trained individually on each of the languages present in the multilingual dataset. The traditional approach has several disadvantages: (i). Complex two-stage architecture, (ii). Failure of LID block leads to the failure of entire system, (iii). Developing monolingual phone recogniser is not feasible for all languages. We propose to use a common multilingual phone-set approach to build Multilingual Phone Recognition System (Multi-PRS).

We address the problem of efficient techniques for speech recognition of code-switched speech. In code-switching, two or more languages are mixed and spoken as if they are one language [15,29]. Code-switching (or code-mixing or language-mixing) involves switching between multiple languages either inter-sententially and intra-sententially [17]. Bilingual code-switching is more common compared to the mixing of more than two languages [26]. The reasons for code-switching include (i) availability of a better word or phrase in another language to express a particular idea, (ii) certain words or phrases are more readily available in the other language, (iii) to show expertise in multiple languages. Code-switching is a common practice across the world in multilingual societies, where a speaker has spoken proficiency in more than one language. In this study, we have considered intra-sentential code-switching between two Indian languages, namely, Kannada (KN) and Urdu (UR), with Kannada sentence being the primary language within which switching occurs to Urdu words and phrases.

The proposed Multi-PRS 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 while also ensuring that the individual language's phones are accurately mapped to the phones in the common phone set. We propose the use of International Phonetic Alphabet (IPA) chart to derive common multilingual phone-set. 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. Few notable works based on common multilingual phone-set approach are reported in [32,33,37,38]. Although there are significant efforts to develop multilingual speech recognizers using Indian languages [2,8,34], not many studies have explored the use of IPA based common multilingual phone-set to develop multilingual phone recognizers using Indian languages. The most recent works on multilingual speech recognition using DNNs are reported in [14,21,22,25,42]. Few notable works on code-switched speech recognition using multilingual speech recognisers are reported in [1,13,16,30,36].

The focus of our work here is to compare two approaches of multilingual speech recognition: (i) one involving using a front-end language-identification stage to detect the language spoken in short intervals of speech and then use

the recognized language's phone recogniser to decode the speech; we refer to this as a LID-switched monolingual approach (LID-Mono), and (ii) using a Multilingual Phone Recognition System based on common multilingual phone-set (Multi-PRS). We further extend the comparison between LID-Mono and Multi-PRS to code-switching scenario. Because the Multi-PRS can seamlessly decode the code-switched speech without regard to the code-switched instances, since the Multi-PRS is designed using a common phone-set between several languages (from which the code-switched speech could switch between any pair of languages) with the corresponding common phone acoustic-models being trained from shared-data from the multiple languages. The rest of the paper is organized as follows: Sect. 2 describes our experimental setup. Section 3 describes and compares the two approaches of multilingual phone recognition. Section 4 extends the comparison to code-switching scenario. Section 5 provides the summary of the paper.

2 Experimental Setup

2.1 Multilingual Speech Corpora

We describe here the details of the speech corpora of the 6 Indian languages used in this work: Kannada (KN), Telugu (TE), Bengali (BN), Odia (OD), Urdu (UR), and Assamese (AS). The speech corpora 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 [12]. Speech corpora contains 16 bit, 16 KHz speech wave files along-with their IPA transcription [39]. The wave files contain read speech sentences of size between 3 to 10 s. Detailed description of the speech corpora is provided in [4,19,23,28]. 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.

Table 1. Statistics of multilingual speech corpora

Language	# Speakers		Duration (in hours)			
	M	F	Train	Dev	Test	Total
Kannada (KN)	7	9	2.80	0.33	0.76	3.89
Telugu (TE)	9	10	4.05	0.47	1.07	5.59
Bengali (BN)	20	30	3.42	0.40	0.99	4.81
Odia (OD)	14	16	3.58	0.36	0.97	4.91
Urdu (UR)	53	6	4.12	0.46	1.04	5.64
Assamese (AS)	8	8	2.39	0.23	0.53	2.39

2.2 Testing Set Speech Corpora for Code-Switching Scenario

We have selected 320 code-switched sentences having Kannada as the primary language and code-switching to Urdu words and phrases. The sentences are carefully chosen to cover all the phonetic units of Kannada and Urdu. *Four* male and *four* female speakers who are bilinguals of Kannada and Urdu are made to read 40 sentences each. The speakers are proficient in both spoken Urdu and Kannada, and tend to produce KN-UR utterances. These sentences were transcribed using IPA symbols and then mapped to the common multilingual phone-set to generate the ground-truth transcription for calculation of PER in the decoding.

Figure 1 shows the distribution of durations of KN and UR languages in code-switched KN/UR test sets. The duration of each utterance range from 3.5 s to 11 s in the data set, within which the Urdu words and phrases occur at durations, from which it can be noted that Kannada segments in an utterance are the longer ones, interspersed with Urdu segments of relatively shorter durations, importantly ranging from 0–0.5 s to 3–3.5 s which typically correspond to short words (<500 ms) and multi-word phrases (of the order of 1–3.5 s). This kind of Urdu segments in the code-switched data plays an important factor in determining how the LID-Mono works, particularly in the choice of the speech interval size on which the front-end LID has to operate.

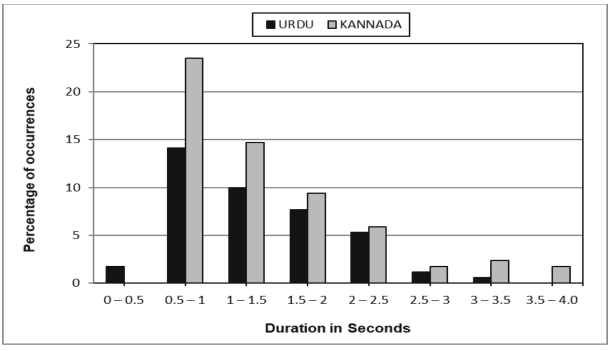

Fig. 1. Distribution of durations of KN and UR languages in code-switched KN/UR test sets.

2.3 Training DNNs

Context dependent 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 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 of DNNs used in the development of Phone Recognition Systems (PRS) are tuned by adjusting the width of the hidden layers. It is found that the DNNs with 5 hidden layers are suitable for building PRSs. 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 [41]. All the experiments are conducted using the open-source speech recognition toolkit - Kaldi [11].

2.4 Extraction of i-vectors

The i-vectors are one of the most widely used features for language recognition. They are fixed dimension feature vectors that are derived from the variable length sequence of front-end features [24]. A DNN is trained for automatic speech recognition using the labelled speech data from Switchboard (SWB1) and Fisher corpora (about 2000 h). Training uses hidden layers with ReLU activation with layer-wise batch normalization. Mel-frequency cepstral coefficients (MFCCs) are extracted from each input utterance and fed to DNNs. The bottleneck features (80 dimension) are extracted from the bottleneck layer of trained DNN [5,6]. The extracted bottleneck features are the front end features. A Gaussian Mixture Universal Background Model (GMM-UBM) is obtained by pooling the front end features from all the utterances in the train dataset. The means of the GMMs are adapted to each utterance using the Baum-Welch statistics of the front-end features. The i-vectors (400 dimension) are computed based on each adapted GMM mean supervector. Since the *SWB1* and *Fisher* corpora used for training the DNNs have the sampling rate of 8 KHz, we have down-sampled the multilingual speech corpora and the testing datasets (see Sects. 2.1 and 2.2) from 16 KHz to 8 KHz for extracting the i-vectors. Detailed description of extraction of i-vectors is given in [7].

3 Approaches for Multilingual Phone Recognition

The following subsections describe the development and comparison of multilingual phone recognizers using two approaches: (i). LID-switched monolingual phone recognition (LID-Mono) approach, and (ii). Multilingual phone recognition using common multilingual phone-set (Multi-PRS) approach.

3.1 LID-switched Monolingual Phone Recognition (LID-Mono) Approach

LID-Mono is a traditional approach for multilingual phone recognition and is shown in Fig. 2. It consists of two stages. In the first stage, the language of the input speech is determined using a language identification block. In the second stage, the input speech utterance is routed to the monolingual phone recognizer of the language identified in stage-1 and the phones present in the input speech are determined. Monolingual phone recognizer is a conventional PRS developed using the data of single language.

We briefly outline the LID system here. There are two approaches for LID, namely, implicit LID and explicit LID. The explicit LID requires phonetic transcription and language models for each language [3,27], whereas the implicit LID does not need either phonetic transcription or language models [10,35]. Since, we do not have language models for the languages considered in this study, we have carried out implicit LID to perform LID. Support Vector Machines (SVM) [20,40] are used to train the LID classifiers. Multi-class SVM is constructed using one-against-one approach (Max-win voting). The radial basis function is used as a kernel. The *LIBSVM* library is used for building SVM models [9]. We have explored both MFCCs and i-vectors as features for building LID systems. The 13-dimensional MFCCs [18] along-with their first and second order derivatives are computed using a frame-length of 25 ms with a frame-shift of 10 ms. The i-vectors are extracted using the procedure described in Sect. 2.4. Table 2 shows the LID accuracy (%) for various language sets using SVMs. Since, the performance of LID using i-vectors outperforms MFCCs, we have considered only the i-vector based LID systems in all our experiments. LID accuracy decreases as the number of languages increase.

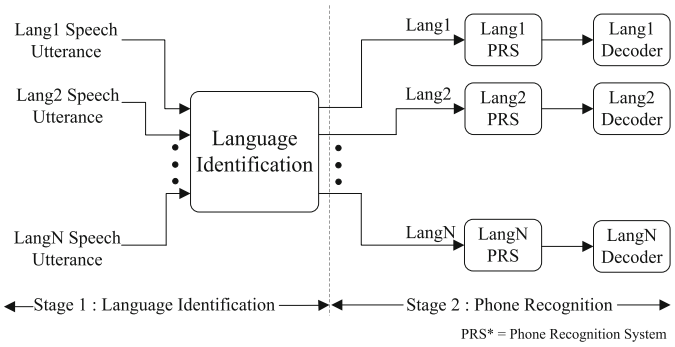

Fig. 2. Multilingual phone recognition using LID-Mono approach.

Table 2. LID accuracy for various language sets using MFCCs and i-vectors.

Languages	LID accuracy (%)	
	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

3.2 Multilingual Phone Recognition Using Common Multilingual Phone-Set Approach (Multi-PRS)

Figure 3 shows the schematic representation of the Multi-PRS. Unlike Fig. 2, which has two stages, Fig. 3 has a single stage - irrespective of any language, Multi-PRS accepts speech input from any language and decodes it into a sequence of phonetic units [31] using a common phone-set.

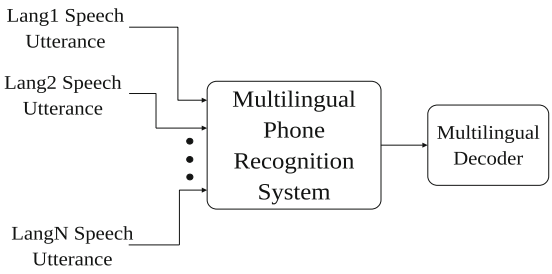

Fig. 3. Multilingual phone recognition using common multilingual phone set approach.

We have developed Multi-PRSs using six Indian languages - KN, TE, BN, OD, UR and AS. The common multilingual phone-set is derived by grouping the acoustically similar IPAs across the 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, 46, 46 phones for 4, 5, and 6 languages, respectively. We have also developed monolingual Phone Recognition Systems (Mono-PRSs) for KN, TE, BN, OD, UR, and AS languages using 36, 35, 34, 36, 35, and 32 phones, respectively. Mono-PRSs are used in second stage of LID-Mono systems as shown in Fig. 2. The Mono-PRSs and Multi-PRSs are trained using CD DNNs.

3.3 Comparison of LID-Mono and Multi-PRS

Phone Error Rate (PER) is determined by comparing the decoded phone labels with the reference transcriptions by performing an optimal string matching using dynamic programming.

Table 3. Phone error rates of multilingual phone recognition systems.

Languages	Approach	
	Multi-PRS	LID-Mono (i-vectors)
KN, BN, OD, UR	31.5	35.5
KN, TE, BN, OD, UR	32.5	35.6
KN, TE, BN, OD, UR, AS	31.9	37.9

Table 3 shows the PERs of LID-Mono and Multi-PRS approaches. It is found that the Multi-PRS systems based on common phone set approach outperform the traditional LID-Mono systems. As the number of languages increase the benefits of Multi-PRSs will be more. Higher the number of languages more the benefit from Multi-PRSs compared to LID-Mono. The use of Multi-PRS has an additional advantage of decoding more number of phones compared to the LID-Mono. This would help the language models to recognise the words more accurately.

We show in Fig. 4, the performance of the LID-Mono and Multi-PRSs on test data drawn from 4, 5 and 6 languages in terms of % LID accuracy and % PER. It can be noted that the LID-Mono has an inherently poor performance marked by decreasing %LID accuracy as the number of language classes increase from 4 to 6, which in turn impacts the % PER to increase in going from 4 to 6 languages. When the LID system makes an error, the LID-switched monolingual phone recognition chooses the wrong language phone acoustic models to decode the input speech and naturally incurs an higher %PER. The %PERs of Multi-PRS

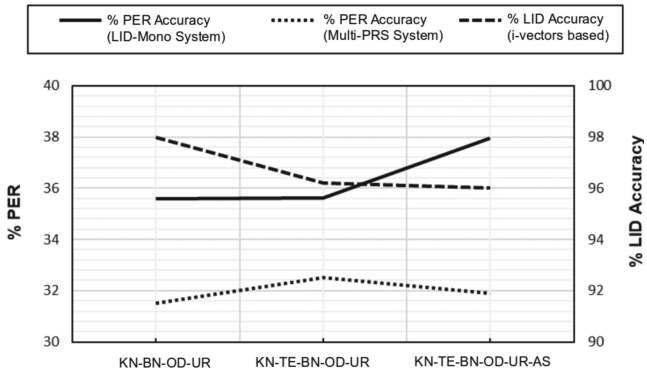

Fig. 4. LID accuracy (%) and PERs (%) of LID-Mono and Multi-PRS approaches.

system, in contrast, have a robust constant performance across the multiple languages clearly arising from its not depending on a front-end LID decision making.

4 LID-Mono and Mutli-PRS Approaches in Code-Switching Application

For comparison of results, in addition to the multilingual phone recognisers based on 4, 5, and 6 languages, we have also developed bilingual phone recognisers using KN and UR languages using both LID-Mono and Multi-PRS approaches. Further, we have analysed how the duration of windows (in seconds) used for performing LID effects the performance of various multilingual phone recognisers. Figure 5 shows a composite display of the performance of LID-Mono and Multi-PRS for different number of test languages, different sizes of the intervals over which the LID makes a decision (called the LID-interval in the x-axis from 500 ms to 5 s and the full utterance), the resulting %LID accuracy and the overall %PER (in the two y-axes).

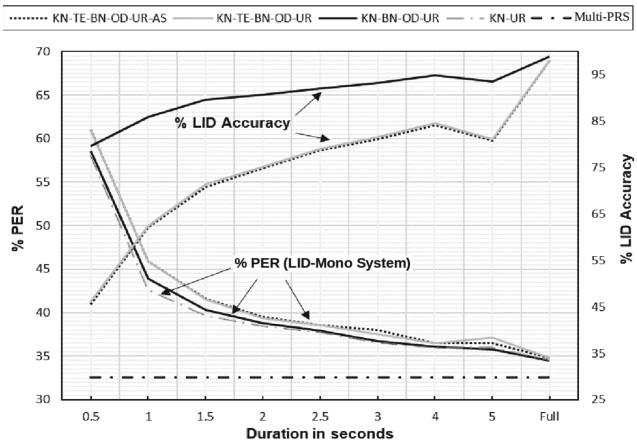

Fig. 5. Comparison of LID-Mono and Multi-PRS systems in code-switching scenario.

Considering the LID-Mono system curves, for small LID-intervals, the LID accuracy is low (45% to 85% for different number of languages considered 6 to 4), with the LID-Mono having to use a wrong language acoustic-model for decoding the corresponding interval. This naturally leads to a very high % PER (of the order of 60%). As the LID-interval size increases the %LID improves, reaching 80–90% for durations of 3 s and 85–95% for longer durations of 5 s (or the full utterance). The corresponding %PER also shows a marked decrease, since the LID-Mono makes a phone-decoding of the given intervals with the correct language acoustic models with increasing accuracy, reaching 40% at 2 secs and down to 35% at 5 secs and more. What is important to note here is that the LID-Mono's performance is dictated by the LID-interval size - smaller

sizes are good for detecting code-switching instances at high time-resolution, but have inherently poor LID accuracies and corresponding poor PER; larger sizes are good for yielding high LID accuracy, with corresponding lower PER, but the code-switching instances are missed due to poor time resolution of the LID decision intervals, i.e. for instance, at a 2 s LID-interval, a good proportion of Urdu segments would have occurred 'within' the 2 s interval (as evident from the code-switching duration distribution in Fig. 1). These segments would then be decoded by the LID decision, which, has no question of being 'correct' since the interval in question is 'mixed' in its ground truth, and the LID has to yield a single language decision, potentially resulting in a mismatch in the language(s) in the 2-sec interval and the single monolingual phone-recognizer that would have been brought into to decode the speech in the 2 s interval. This problem becomes more acute as the LID-interval size increases.

On the contrary, the KN-UR Multi-PRS using common multilingual phone-set approach offers a consistent performance of 32.7% without having any LID-interval in its pipe-line, and hence is robust to arbitrary code-switching durational distributions (as in Fig. 1). This makes the Multi-PRS the natural choice to recognize code-switched speech, with practically no particular merit to choose the LID-Mono system, which suffers from the trade-off discussed above, higher design complexity of having to design a LID system (to recognize multiple language classes), and having to design multiple monolingual phone recognition systems.

5 Conclusions

We have developed and compared LID-Mono and Multi-PRS approaches of multilingual phone recognition. We have extended the same study to code-switched speech recognition scenario using code-switched utterance of two Indian languages (Kannada and Urdu). We have studied the performance characteristics of LID-Mono and Multi-PRS approaches with respect to several underlying parameters, such as the interval over which the LID makes a decision and the number of languages on which the LID is designed for the LID-Mono systems, the means of arriving at a common phone-set for the Multi-PRS and shown that while the LID-Mono system suffers from inherent trade-off's between interval sizes, the Multi-PRS offers a robust performance for arbitrary code-switching data.

References

1. Vu, N.T., et al.: A first speech recognition system for Mandarin-English code-switch conversational speech. In: ICASSP, pp. 4889–4892 (2012)
2. Mohan, A., Rose, R., Ghalehjegh, S.H., Umesh, S.: Acoustic modelling for speech recognition in Indian languages inan agricultural commodities task domain. Speech Commun. **56**, 167–180 (2014)
3. Sai Jayram, A.K.V., Ramasubramanian, V., Sreenivas, T.V.: Language identification using parallel sub-word recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSAP), vol. 1 (2003)

4. Sarma, B.D., Sarma, M., Sarma, M., Prasanna, S.R.M.: Development of Assamese phonetic engine: some issues. In: IEEE INDICON, pp. 1–6 (2013)
5. Jiang, B., Song, Y., Wei, S., Liu, J.H., McLoughlin, I., Dai, L.: Deep bottleneck features for spoken language identification. PLoS ONE **9**(7), e100795 (2014)
6. Jiang, B., Song, Y., Wei, S., Wang, M., McLoughlin, I., Dai, L.: Performance evaluation of deep bottleneck features for spoken language identification. In: International Symposium on Chinese Spoken Language Processing, pp. 143–147 (2014)
7. Padi, B., Ramoji, S., Yeruva, V., Kumar, S., Ganapathy, S.: The LEAP language recognition system for LRE 2017 challenge - improvements and error analysis. In: The Speaker and Language Recognition Workshop, Odyssey (2018)
8. Kumar, C.S., Mohandas, V.P., Haizhou, L.: Multilingual speech recognition: a unified approach. In: INTERSPEECH, pp. 3357–3360 (2005)
9. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011). software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
10. Nandi, D., Pati, D., Rao, K.S.: Implicit processing of LP residual for language identification. Comput. Speech Lang. **41**, 68–87 (2017)
11. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE Workshop on ASRU (2011). http://kaldi-asr.org/
12. Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian Languages. http://speech.iiit.ac.in/svldownloads/pro_po_en_report/
13. Yilmaz, E., Heuvel, H.V.D., Leeuwen, D.V.: Code-switching detection using multilingual DNNs. In: IEEE Workshop on Spoken Language Technology for Under-Resourced Languages, pp. 159–166 (2016)
14. Heigold, G., et al.: Multilingual acoustic models using distributed deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2013)
15. Kroll, J.F., De Groot, A.M.B. (eds.): Handbook of Bilingualism: Psycholinguistic Approaches. Oxford University Press, New York (2005)
16. Bhuvanagirir, K., Kopparapu, S.K.: Mixed language speech recognition without explicit identification of language. Am. J. Sig. Process. **2**(5), 92–97 (2012)
17. Jorschick, L., Quick, A.E., Glasser, D., Lieven, E., Tomasello, M.: German-English-speaking children-s mixed NPs with 'correct' agreement. Bilingualism: Language and Cognition **14**(2), 173–183 (2011)
18. Rabiner, L., Juang, B., Yegnanarayana, B.: Fundamentals of Speech Recognition. Pearson Education, New Delhi (2008)
19. Madhavi, M.C., Sharma, S., Patil, H.A.: Development of language resources for speech application in Gujarati and Marathi. In: IEEE International Conference on Asian Language Processing (IALP), vol. 1, pp. 115–118 (2014)
20. Li, M., Suo, H., Wu, X., Lu, P., Yan, Y.: Spoken language identification using score vector modeling and support vector machine. In: Interspeech, pp. 350–353 (2007)
21. Muller, M., Waibel, A.: Using language adaptive deep neural networks for improved multilingual speech recognition. In: International Workshop on Spoken Language Translation (IWSLT) (2015)
22. Muller, M., Stuker, S., Waibel, A.: Towards improving low-resource speech recognition using articulatory and language features. In: International Workshop on Spoken Language Translation (IWSLT), pp. 1–7 (2016)
23. Shridhara, M.V., Banahatti, B.K., Narthan, L., Karjigi, V., Kumaraswamy, R.: Development of Kannada speech corpus for prosodically guided phonetic search engine. In: Sixteenth International Oriental COCOSDA (2013)

24. Dehak, N., Carrasquillo, P.A.T., Reynolds, D., Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
25. Vu, N.T., et al.: Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: ICASSP (2014)
26. Heredia, R.R., Altarriba, J.: Bilingual language mixing: why do bilinguals code-switch? Curr. Dir. Psychol. Sci. **10**, 164–168 (2001)
27. Santosh Kumar, S.A., Ramasubramanian, V.: Automatic language identification using ergodic-HMM. In: ICASSP, pp. 609–612 (2005)
28. Kumar, S.B.S., Rao, K.S., Pati, D.: Phonetic and prosodically rich transcribed speech corpus in Indian languages: Bengali and Odia. In: O-COCOSDA, pp. 1–5 (2013)
29. Ford, S.: Language mixing among bilingual children. http://www2.hawaii.edu/~sford/research/mixing.htm
30. Kim, S., Seltzer, M.L.: Towards language-universal end-to-end speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4914–4918 (2018)
31. Siniscalchi, S.M., Lyu, D., Svendsen, T., Lee, C.: Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. IEEE Trans. Acoust. Speech Signal Process. **20**(3), 875–887 (2012)
32. Stuker, S., Schultz, T., Metze, F., Waibel, A.: Multilingual articulatory features. In: ICASSP, vol. 1, pp. 144–147 (2003)
33. Stuker, S., Metze, F., Schultz, T., Waibel, A.: Integrating multilingual articulatory features into speech recognition. In: INTERSPEECH, pp. 1033–1036 (2003)
34. Gangashetty, S.V., Sekhar, C.C., Yegnanarayana, B.: Spotting multilingual consonant-vowel units of speech using neural network models. In: Faundez-Zanuy, M., Janer, L., Esposito, A., Satue-Villar, A., Roure, J., Espinosa-Duro, V. (eds.) NOLISP 2005. LNCS (LNAI), vol. 3817, pp. 303–317. Springer, Heidelberg (2006). https://doi.org/10.1007/11613107_27
35. Nagarajan, T., Murthy, H.A.: A pair-wise multiple codebook approach to implicit language identification. In: Workshop on Spoken Language Processing, pp. pp. 101–108 (2003)
36. Schultz, T.: Multilingual automatic speech recognition for code-switching speech. In: The 9th International Symposium on Chinese Spoken Language Processing (2014)
37. Schultz, T., Waibel, A.: Language independent and language adaptive acoustic modeling for speech recognition. Speech Commun. **35**, 31–51 (2001)
38. Schultz, T., Kirchhoff, K.: Multilingual Speech Processing. Academic Press, Amsterdam (2006)
39. The International Phonetic Association: Handbook of the International Phonetic Association. Cambridge University Press (2007). https://www.internationalphoneticassociation.org/
40. Campbell, W.M., Singer, E., Torres-Carrasquillo, P.A., Reynolds, D.A.: Language recognition with support vector machines. In: Proceedings of the Odyssey: The Speaker and Language Recognition Workshop, pp. 41–44 (2004)
41. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: ICASSP, pp. 215–219 (2014)
42. Miao, Y., Metze, F.: Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training. In: INTERSPEECH, pp. 2237–2241 (2013)