INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multilingual Speech Recognition Using Language-Specific Phoneme Recognition as Auxiliary Task for Indian Languages

Hardik B. Sailor[1]*, *Thomas Hain*[2]

[1]Samsung Research Institute, Bangalore, India
[2]Speech and Hearing Research Group, The University of Sheffield, UK

h.sailor@samsung.com, t.hain@sheffield.ac.uk

Abstract

This paper proposes a multilingual acoustic modeling approach for Indian languages using a Multitask Learning (MTL) framework. Language-specific phoneme recognition is explored as an auxiliary task in MTL framework along with the primary task of multilingual senone classification. This auxiliary task regularizes the primary task with both the context-independent phonemes and language identities induced by language-specific phoneme. The MTL network is also extended by structuring the primary and auxiliary task outputs in the form of a Structured Output Layer (SOL) such that both depend on each other. The experiments are performed using a database of the three Indian languages Gujarati, Tamil, and Telugu. The experimental results show that the proposed MTL-SOL framework performed well compared to baseline monolingual systems with a relative reduction of 3.1-4.4 and 2.9-4.1 % in word error rate for the development and evaluation sets, respectively.

Index Terms: Multilingual, Auxiliary task learning, Structured Output Layer (SOL)

1. Introduction

Recently, there is a significant interest in developing multilingual speech and language technologies. This is of particular importance for countries with many languages such as India, Russia, or South Africa. India has 22 official languages with an additional 1500 minor languages/dialects. Apart from a few major languages, most of the languages are low resourced. This poses several challenges to develop speech technologies such as Automatic Speech Recognition (ASR). Multilingual speech recognition using Deep Neural Networks (DNN) is a promising research direction towards building an ASR system for low resource languages. One of the limitations of the multilingual acoustic model is that sometimes it precludes the fine-tuning aspects of the particular language [1]. To mitigate such a problem, Multitask Learning (MTL) has emerged compared to the traditional single-task learning approach [2]. In the deep learning era, the hard parameter sharing in MTL involves learning parallel tasks with the shared hidden layers and the task-specific hidden layers. With multilingual MTL approaches, it is possible to train language-aware ASR tasks or language identification as an auxiliary task [3], [4].

The goal of an auxiliary task in MTL is to enable the model to learn feature representations that are beneficial to the primary task. However, it is important to choose related tasks in MTL otherwise the performance of the primary task deteriorates (called as a negative transfer) [2]. In this paper, MTL with an auxiliary task is proposed to improve multilingual ASR for the three Indian languages. Language-specific phoneme recognition is proposed as the auxiliary task since it is more related to the primary task of multilingual senones classification in DNN. Our key contributions in this paper are as follows:

- The MTL approach of monophone regularization proposed in [5] is extended using language-specific phoneme recognition as an auxiliary task for multilingual acoustic modeling.
- The Structured Output Layer (SOL) is applied in the MTL model to improve the proposed auxiliary task learning. Here, the SOL is used for both the primary and auxiliary tasks compared to the earlier approach of the SOL in the primary task [6].
- ASR experiments were performed using the three Indian languages data released during Interspeech 2018 leading to competitive performance.

2. Related Work

The literature of multilingual/cross-lingual ASR is well summarized in a series of survey papers [7–9]. Here, all approaches that are more related to the proposed framework for Indian languages are described. To generate alignments for DNN training, the GMM-HMM systems are initially trained either using a Universal Phone Set (UPS) or using the Union of Phoneme Set (UoPS) [10], [11]. The UPS-based model is trained using a single DNN with softmax representing multilingual senone labels [12].

In the Shared Hidden Layer Multilingual DNN (SHL-MDNN), the hidden layers (except the last hidden layer) are shared across languages and the output layers are language-specific [13], [14]. Many studies used bottleneck features (BNF) [15] and Language Feature Vectors (LFV) for language adaptation [16]. The SHL-MDNN is trained using an MTL approach [2]. The goal of MTL in the SHL-MDNN is to generalize the performance of each language task by sharing network parameters. Adversarial training was also proposed in the context of MTL for a multilingual ASR [17]. There are several studies published in the ASR domain that exploit auxiliary task-based MTL. In the ASR literature, the auxiliary task includes classification of speakers [18], monophones [5], dialects/accents [19], or language identification [17] in the multilingual case. Cross-entropy regularization was also proposed as an auxiliary task to improve the performance of a sequence-level ASR training [20]. Language-independent multilingual End-to-end ASR approach was proposed to jointly identify the language and the character set of a particular language [21]. In [6], the Structured Output Layer (SOL) approach was proposed such that senone outputs are dependent on the monophone hidden layer features.

[*] Work was done when author was at University of Sheffield, UK.

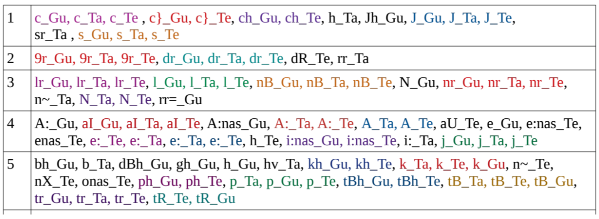
http://dx.doi.org/10.21437/Interspeech.2020-2739

In Interspeech 2018, a low resource speech recognition challenge in Indian languages was organized [22]. The top-performing systems in the challenge used UPS for either a hybrid DNN-HMM [23], [24], or an end-to-end ASR approach [25]. Other notable multilingual approaches from the challenge include articulatory features [26] and a joint acoustic model using Subspace Gaussian mixture models (SGMM) and end-to-end ASR [27]. Apart from the challenge in Interspeech 2018, there are various approaches proposed for ASR in Indian languages [28–33]. Recently, a study in [34] proposed a transformer-based end-to-end multilingual model for Indian languages that also includes language information as a one-hot vector and embeddings.

3. Multilingual Multi-task Learning

3.1. Multilingual Acoustic Modeling

Acoustic modeling is performed using a DNN-HMM system where the alignments are generated as a part of GMM-HMM training. The multilingual UoPS set is created using language tags attached to the phonemes of a particular language. The monophone and triphone GMM-HMM systems are trained by merging speech data of the three languages. A top-down binary clustering of the data is applied to generate phonetic questions using the clustering method proposed in [35]. This approach is similar to [10] where the language tags are attached to phonemes. However, the clustering technique in [10] also includes specific questions about the language and its phonetic categories. Here, a question set is automatically generated using a clustering technique in the Kaldi toolkit [35]. In the decision-tree clustering process, a question represents a set of phones that have shared roots. Examples of multilingual phonemes clustered in the Kaldi toolkit is shown in Figure 1. It can be observed that the clustering process automatically combines phonemes that are similar in either all three or any of the two languages. The triphone alignments are further refined using an LDA-MLLT system as suggested in [36]. The senone alignments are then used in multilingual DNN training as shown in Figure 2 (a). The acoustic modeling is performed using a Bidirectional Gated Recurrent Units (BiGRU) [37].



1	c_Gu, c_Ta, c_Te , c}_Gu, c}_Te, ch_Gu, ch_Te, h_Ta, Jh_Gu, J_Gu, J_Ta, J_Te, sr_Ta , s_Gu, s_Ta, s_Te
2	9r_Gu, 9r_Ta, 9r_Te, dr_Gu, dr_Ta, dr_Te, dR_Te, rr_Ta
3	lr_Gu, lr_Ta, lr_Te, l_Gu, l_Ta, l_Te, nB_Gu, nB_Ta, nB_Te, N_Gu, nr_Gu, nr_Ta, nr_Te, n~_Ta, N_Ta, N_Te, rr=_Gu
4	A:_Gu, aI_Gu, aI_Ta, aI_Te, A:nas_Gu, A:_Ta, A:_Te, A_Ta, A_Te, aU_Te, e_Gu, e:nas_Te, enas_Te, e:_Te, e:_Ta, e:_Ta, e:_Te, h_Te, i:nas_Gu, i:nas_Te, i:_Ta, j_Gu, j_Ta, j_Te
5	bh_Gu, b_Ta, dBh_Gu, gh_Gu, h_Gu, hv_Ta, kh_Gu, kh_Te, k_Ta, k_Te, k_Gu, n~_Te, nX_Te, onas_Te, ph_Gu, ph_Te, p_Ta, p_Gu, p_Te, tBh_Gu, tBh_Te, tB_Ta, tB_Te, tB_Gu, tr_Gu, tr_Ta, tr_Te, tR_Te, tR_Gu

Figure 1: *List of clustered phones generated in Kaldi [35]. Here, each phonemes is attached a tag _Gu, _Ta, or _Te for Gujarati, Tamil and Telugu, respectively. Best viewed in color.*

3.2. Auxiliary Task Learning

Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_T\}$, $\mathbf{x}_t \in \mathbb{R}^D$, $t = 1, ..., T$ is a sequence of speech frames of D-dimensional feature vectors. Let $\mathbf{Y} = \{y_1, ..., y_T\}$, $y_t \in \mathbb{R}$ and $\mathbf{Z} = \{z_1, ..., z_T\}$, $z_t \in \mathbb{R}$ are the sequences of senones and phoneme labels (i.e., monophone targets) aligned with $\mathbf{X}$. Here, phoneme labels also include language tags compared to the senone labels where phonemes from different languages (along with tags) are clustered in the triphone model. The shared DNN layers in the MTL framework can be observed as a feature extractor network. The schematic diagram of the proposed MTL system is shown in Figure 2 (b). With θ_f, θ_s, and θ_a are being the parameters of feature extractor, senones, and auxiliary task classifier, respectively then the loss functions can be written as:

$$\mathcal{L}_{senone}(\theta_f, \theta_s) = -\frac{1}{T}\sum_{t=1}^{T} \log p(y_t|\mathbf{x}_t; \theta_f, \theta_s) \quad (1)$$

$$\mathcal{L}_{auxiliary}(\theta_f, \theta_a) = -\frac{1}{T}\sum_{t=1}^{T} \log p(z_t|\mathbf{x}_t; \theta_f, \theta_a) \quad (2)$$

The final loss of the MTL framework is given as,

$$\mathcal{L}_{total} = \mathcal{L}_{senone} + \lambda\mathcal{L}_{auxiliary} \quad (3)$$

where $\lambda \in [0, 1]$ is a task weighting parameter. Based on this formulation, the role of this auxiliary task is to generalize the primary task by providing a regularization. Our proposed auxiliary task of classifying monophone targets with language tags gives additional information about the language of a particular frame. This auxiliary task is different than a language identification (LID) task where the labels are language identities only. Hence, the primary task is regularized using both the phoneme and language information indirectly obtained given the phonemes of a particular language. We also experimentally confirmed that the proposed auxiliary task is more beneficial than the LID task.

To exploit the task relationships, we investigate the Structured Output Layer (SOL) approach in the context of MTL [6]. In this work, we extend the effectiveness of the SOL approach for both the primary and auxiliary tasks compared with [6] where it was proposed for the primary task only. The SOL is added in the network as shown in Figure 2 (c). Here, $\mathbf{a}_y$ and $\mathbf{a}_z$ are the activations of task-specific BiGRU layers of senone and phoneme task, respectively. The SOL is added for both the tasks as follows:

$$p(y_t|\mathbf{x}_t) = \text{softmax}(\mathbf{y}_{SOL}) \text{ where } \mathbf{y}_{SOL} = \sigma(\mathbf{a}_z) + \mathbf{a}_y \quad (4)$$

$$p(z_t|\mathbf{x}_t) = \text{softmax}(\mathbf{z}_{SOL}) \text{ where } \mathbf{z}_{SOL} = \sigma(\mathbf{a}_y) + \mathbf{a}_z \quad (5)$$

Here, the sigmoid $\sigma(\cdot)$ hidden layer is used in the SOL where the parameters of a network are shared between both the tasks. Other activation functions were also tried instead of a sigmoid; however, no performance gains were observed (results not shown here). Since the sigmoid activation compresses the information between 0 and 1, it is well suited for feature augmentation so that the primary task features do not change significantly. The features from the auxiliary task are augmented with the BiGRU response of the primary task and vice-versa. The difference between MTL and MTL-SOL is that in the case of MTL, the auxiliary task network is discarded during final testing. However, a part of the auxiliary network is still active (except the softmax layer) in the case of MTL-SOL.

4. Experimental Setup

4.1. Database

All experiments are performed with a database of Indian languages released as a part of the “Low Resource Speech Recognition Challenge for Indian Languages” during Interspeech 2018. The database contains three languages, namely, Gujarati,

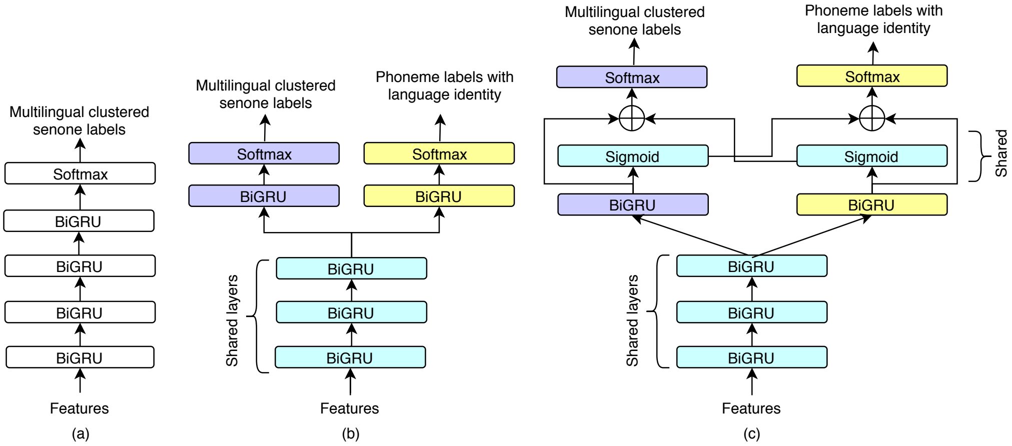


Figure 2: *(a) Multilingual, (b) MTL, and (c) MTL-SOL systems. The blue color indicates shared layers.*

Tamil, and Telugu. The statistics of a database including training, development, and evaluation set are shown in Table 1. The lexicon is provided by the organizers created using the CMU Festvox Indic frontend phoneme set.

Table 1: *Database statistics of Indian languages [22].*

Lang.	**# Words**	**# Phonemes**	**Duration in hrs**		
			Train	**Dev**	**Eval**
Tamil	57883	38	40	5	4.2
Telugu	48686	56	40	5	4.2
Gujarati	41238	54	40	5	5

4.2. ASR System Building

The feature extraction, language modeling, and GMM-HMM training are performed in the Kaldi toolkit [35]. The GMM-HMM systems are trained using 39-D Mel Frequency Cepstral Coefficients (MFCC). The number of senones for each system are varied and decided using recognition performance on the development set. The optimal number of triphones is 3500 and 7000 for monolingual and multilingual systems, respectively. Triphone alignments are further refined using the LDA-MLLT system. The DNN acoustic models are built in the PyTorch-Kaldi toolkit [38]. The DNN systems are trained using 40-D log Mel filterbank features. The baseline monolingual and proposed multilingual systems are trained using a deep BiGRU network with 4 layers and 650 hidden units each. The MTL system has 3 shared BiGRU layers and one task-specific layer, each with 650 hidden units. The BiGRU networks are trained for 24 epochs with the Adam algorithm. The decoding is done using a language-specific 3-gram LM.

5. Experimental Results

The performance of UoPS is better than UPS for GMM-HMM systems with a relative WER reduction of 1-3.49 % compared with monolingual systems. The detailed GMM-HMM results are omitted due to space limitations. Hence, the UoPS multilingual system is used to generate alignments for DNN-HMM training. The experimental results of the DNN-HMM systems are reported in Table 2 for the development set. The multilingual system reduced the WER with a relative WER reduction of 1.97, 2.18, and 0.79 % for Gujarati, Telugu, and Tamil language, respectively. These results show the significance of multilingual DNN training using UoPS for acoustic modeling.

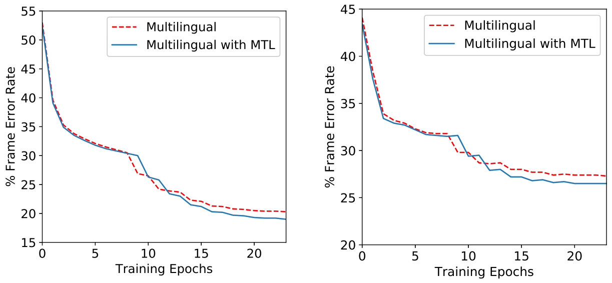
The results of MTL approaches are also shown in Table 2 where $\lambda = 1$. The multilingual MTL system with language identification (LID) as an auxiliary task did not perform well compared with a multilingual system without MTL. We have also experimented with different values of λ in eq. (3). The lower values of λ reduce WER with lower bound on the multilingual ASR system (i.e., $\lambda = 0$).

The primary reason for this poor performance could be an unrelated task in the MTL framework. It is shown that when auxiliary tasks are not related to a primary task the learned features in shared layers can be influenced by outlier features from the auxiliary tasks that may degrade the performance of a primary task [2]. Hence, the proposed auxiliary task of a phoneme recognition with language tags is more suitable in the MTL framework. It improves the system performance significantly compared with the monolingual baseline for all three languages with a relative WER reduction of 3.57, 4.2, and 3.02 % for Gujarati, Telugu, and Tamil, respectively. Different values of λ in eq. (3) were tried in this paper for the proposed auxiliary task; however, the best performance is obtained using $\lambda = 1$.

Table 2: *Experimental results of monolingual and multilingual ASR systems on the development set in % WER.*

Method	Gujarati	Telugu	Tamil
Monolingual	13.16	19.29	16.91
Multilingual	12.90	18.87	16.79
Proposed MTL with phoneme recog.	**12.69**	**18.48**	**16.40**
MTL with language identification	13.44	19.14	17.11

Compared with the multilingual single task system, the MTL system with phoneme recognition gave a relative reduction of 1.63, 2.07, and 2.32 % in WER for Gujarati, Telugu, and Tamil, respectively. The reason for this WER reduction in the primary task is due to task relatedness in the MTL framework. The auxiliary task reduces the senones Frame Error Rate (FER) during DNN training. The FER for training and cross-validation (CV) set is shown in Figure 3 for all the training epochs. It can be observed that the MTL system converges better at a lower FER at the end of training compared with the multilingual system without using the auxiliary task approach.



(a) Training Set

(b) Cross-validation Set

Figure 3: *% FER across training epochs.*

The results of using an MTL-SOL system are shown in Table 3 for the development set. The MTL-SOL system performs slightly better than the MTL system for Gujarati and Tamil languages with a relative WER reduction of 4.41 and 3.13 %, respectively. However, it did not give improvements to the Telugu language. One of the reasons for this could be a highly optimized primary task for Telugu language and hence adding the SOL could not help further. The addition of extra constraints and regularization on the SOL will be left for future studies.

Table 3: *Results in % WER (and relative WER reduction) using SOL in MTL framework on the development set.*

Method	Gujarati	Telugu	Tamil
Monolingual	13.16	19.29	16.91
MTL	12.69 (3.57)	18.48 (4.2)	16.40 (3.02)
MTL-SOL	12.58 (4.41)	18.48 (4.2)	16.38 (3.13)

Results of the evaluation sets are shown in Table 4. The multilingual system performed well for Gujarati and Telugu, however, it did not give improvements for the Tamil language. Our proposed MTL-SOL system gave a relative WER reduction of 3.57, 4.16, and 2.86 for Gujarati, Telugu, and Tamil, respectively compared with a monolingual system. It also gave a relative WER reduction of 0.54, 3.22, and 2.97 for Gujarati, Telugu, and Tamil, respectively compared with a multilingual system.

The statistical significance of the relative WER reduction is justified using a bootstrap technique proposed in [39]. The statistical significance tests were conducted using a compute-wer-bootci tool in the Kaldi toolkit [35]. Table 4 also shows the % Probability of Improvement (POI) measure estimated using bootstrap samples for the proposed system compared with the baseline multilingual ASR. Higher % POI value means statistically significant improvement with the maximum value being 100 and the minimum value being 0. Table 4 shows a significant improvement using a proposed approach compared with the baseline for all the three languages. Relatively lower % POI for the Gujarati language (81.53) is expected due to a small relative improvement of 0.54 % using the MTL-SOL system compared with the multilingual system (Table 4).

Comparison with the challenge baseline and top-performing systems are given in Table 5. The multilingual models of this paper performed well compared with the monolingual models from the challenge baseline. The top systems used data augmentation and fine-tuning a DNN model for a language after the multilingual training. Also, [23] and [24] used discriminative LF-MMI training criteria. The best performing system [23] also utilized the recurrent neural network language models. Compared with these systems, the proposed approach in this paper did not use any data augmentation and a two-stage approach of fine-tuning on a particular language. Our goal in this paper is to show the significance of the proposed auxiliary task of language-specific phoneme recognition and the MTL framework.

Table 4: *Results on the evaluation set in % WER. The numbers in the parentheses shows % POI at 95 % confidence interval.*

Method	Gujarati	Telugu	Tamil
Monolingual	19.05	19.45	16.80
Multilingual	18.47	19.26	16.82
MTL-SOL	**18.37** (81.53)	**18.64** (99.94)	**16.32** (100)

Table 5: *Comparison with other approaches.*

Method	Gujarati	Telugu	Tamil
Proposed MTL-SOL model	18.37	18.64	16.32
Challenge baseline [22]	20.0	21.0	19.5
BUT [23]	14.06	14.71	13.92
Cogknit [24]	17.69	17.14	16.07
CSALT-LEAP/USC [25]	19.31	17.59	16.32

6. Summary and Conclusions

The multilingual ASR system using the MTL framework was presented for the Indian languages. Language-specific phoneme recognition as an auxiliary task in MTL was investigated. The proposed MTL framework reduced % FER in the primary task of senone classification and hence achieved lower % WER compared with monolingual systems. The performance was slightly further improved by the introduction of SOL for primary and auxiliary tasks in MTL. The experiments on the three Indian languages showed that our proposed MTL-SOL system performed well compared with the baseline monolingual and multilingual systems. Future work includes investigation of MTL with the attention-based architecture for multilingual ASR [40].

7. Acknowledgements

We sincerely thank VoiceBase Inc. for research funding and University of Sheffield for research progress. We acknowledge SpeechOcean.com and Microsoft for ASR Challenge data.

8. References

[1] H. Bourlard, J. Dines, M. Magimai-Doss, P. N. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, and F. Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, no. 5, pp. 885–915, 2011.

[2] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[3] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8619–8623.

[4] D. Chen and B. K. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1172–1183, 2015.

[5] P. Bell, P. Swietojanski, and S. Renals, "Multitask learning of context-dependent targets in deep neural network acoustic models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 238–247, Feb 2017.

[6] P. Swietojanski, P. Bell, and S. Renals, "Structured output layer with auxiliary targets for context-dependent acoustic modelling," in *Interspeech*, 2015.

[7] P. Fung and T. Schultz, "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 89–97, May 2008.

[8] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," vol. 56. Elsevier, 2014, pp. 85–100.

[9] H. Sailor, A. Patil, and H. Patil, "Advances in Low Resource ASR: A Deep Learning Perspective," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 2018, pp. 15–19.

[10] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31 – 51, 2001.

[11] B. Imperl, Z. Kačič, B. Horvat, and A. Žgank, "Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones," *Speech Communication*, vol. 39, no. 3, pp. 353 – 366, 2003.

[12] 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 *IEEE ICASSP*, May 2014, pp. 7639–7643.

[13] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE ICASSP*, May 2013, pp. 7304–7308.

[14] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*, Vancouver, BC, Canada, May 2013, pp. 7319–7323.

[15] F. Grézl, E. Egorova, and M. Karafiát, "Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure," in *IEEE SLT*, 2014, pp. 48–53.

[16] M. Müller, S. Stüker, and A. Waibel, "Neural codes to factor language in multilingual speech recognition," in *IEEE ICASSP*, May 2019, pp. 8638–8642.

[17] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 3, pp. 621–630, March 2019.

[18] G. Pironkov, S. Dupont, and T. Dutoit, "Speaker-aware long short-term memory multi-task learning for speech recognition," in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug 2016, pp. 1911–1915.

[19] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *ICASSP*, April 2018, pp. 1–5.

[20] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI." in *Interspeech*, 2016, pp. 2751–2755.

[21] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *ASRU*, Dec 2017, pp. 265–271.

[22] B. M. L. Srivastava, S. Sitaram *et al.*, "Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages," in *Proc. The 6th Intl. Workshop on SLTU*, 2018, pp. 11–14.

[23] B. Pulugundla, M. K. Baskar, S. Kesiraju, E. Egorova, M. Karafiát, L. Burget, and J. Černocký, "BUT system for low resource Indian language ASR," in *ISCA Interspeech*, 2018, pp. 3182–3186.

[24] N. Fathima, T. Patel, M. C, and A. Iyengar, "TDNN-based multilingual speech recognition system for low resource Indian languages," in *Interspeech*, 2018, pp. 3197–3201.

[25] J. Billa, "ISI ASR system for the low resource speech recognition challenge for Indian languages," in *Interspeech*, 2018, pp. 3207–3211.

[26] V. M. Shetty, R. A. Sharon, B. Abraham, T. Seeram, A. Prakash, N. Ravi, and S. Umesh, "Articulatory and stacked bottleneck features for low resource speech recognition," in *Interspeech*, 2018, pp. 3202–3206.

[27] H. Krishna *et al.*, "An exploration towards joint acoustic modeling for Indian languages: IIIT-H submission for low resource speech recognition challenge for Indian languages," in *Interspeech*, 2018, pp. 3192–3196.

[28] C. S. Kumar, V. Mohandas, and H. Li, "Multilingual speech recognition: A unified approach," in *INTERSPEECH*, 2005.

[29] M. J. F. Gales, K. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED," in *SLTU*, 2014.

[30] M. K E *et al.*, "Indian languages ASR: A multilingual phone recognition framework with IPA based common phone-set, predicted articulatory features and feature fusion," in *Interspeech*, 2018, pp. 1016–1020.

[31] D. Dash, M. Kim, K. Teplansky, and J. Wang, "Automatic speech recognition with articulatory information and a unified dictionary for Hindi, Marathi, Bengali and Oriya," in *Interspeech*, 2018, pp. 1046–1050.

[32] A. Kannan *et al.*, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," in *Interspeech*, 2019, pp. 2130–2134.

[33] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4904–4908.

[34] V. M. Shetty, M. Sagaya Mary N J, and S. Umesh, "Improving the performance of transformer based low resource speech recognition for Indian languages," in *IEEE ICASSP*, 2020, pp. 8279–8283.

[35] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on ASRU, Big Island, Hawaii, USA*, 2011, pp. 1–4.

[36] S. P. Rath, D. Povey, K. Veselý, and J. ernocký, "Improved feature processing for deep neural networks," in *INTERSPEECH*, 2013.

[37] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*, Oct. 2014, pp. 1724–1734.

[38] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *IEEE ICASSP*, 2019.

[39] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *IEEE ICASSP*, vol. 1, 2004, pp. I–409.

[40] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1871–1880.