Interspeech 2018
2-6 September 2018, Hyderabad

An Exploration Towards Joint Acoustic Modeling for Indian Languages: IIIT-H submission for Low Resource Speech Recognition Challenge for Indian languages, INTERSPEECH 2018

Hari Krishna Vydana, Krishna Gurugubelli, V V V Raju, Anil Kumar Vuppala

Speech Processing Laboratory, KCIS
International Institute of Information Technology, Hyderabad, India
{hari.vydana,krishna.gurugubelli,vishnu.raju}@research.iiit.ac.in,
anil.vuppala@iiit.ac.in

Abstract

India being a multilingual society, a multilingual automatic speech recognition system (ASR) is widely appreciated. Despite different orthographies, Indian languages share same phonetic space. To exploit this property, a joint acoustic model has been trained for developing multilingual ASR system using a common phone-set. Three Indian languages namely Telugu, Tamil and, Gujarati are considered for the study. This work studies the amenability of two different acoustic modeling approaches for training a joint acoustic model using common phone-set. Sub-space Gaussian mixture models (SGMM), and recurrent neural networks (RNN) trained with connectionst temporal classification (CTC) objective function are explored for training joint acoustic models. From the experimental results, it can be observed that the joint acoustic models trained with RNN-CTC have performed better than SGMM system even on 120 hours of data (approx 40 hrs per language). The joint acoustic model trained with RNN-CTC has performed better than monolingual models, due to an efficient data sharing across the languages. Conditioning the joint model with language identity had a minimal advantage. Sub-sampling the features by a factor of 2 while training RNN-CTC models has reduced the training times and has performed better.

Index Terms: Speech recognition, Joint acoustic model, low-resource, common phone set, Indian languages, RNN-CTC, SGMM

1. Introduction

Though multilingual automatic speech recognition (ASR) systems are widely appreciated in India. Minimal attempts have been made due to the scarcity of resources required for developing state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems [1, 2, 3, 4, 5]. Resources required for developing an ASR system can be broadly grouped to two aspects i.e., transcribed data and pronunciation models. As Indian languages are syllabic in nature, the pronunciation models could be generated from a simple rule-based parser [6, 7, 8, 9]. Low resource in an Indian scenario majorly reflects lack of transcribed data. Indian languages have certain advantageous properties such as sharing same phonetic space and differing in phonotactics [6]. Indian languages differ in prosody i.e., duration, intonation, and prominence associated with a syllable [6]. These properties could be beneficial for developing multi-lingual ASR systems, but the selection of an appropriate phone-set and the suitable acoustic modeling approach are crucial for achieving better performances. Thus in this study, we explore two different acoustic modeling approaches for training a joint acoustic model for Indian languages.

Traditional speech recognition systems use hidden Markov model-Gaussian mixture models (HMM-GMM) as acoustic models. In a HMM-GMM acoustic model, HMMs model the tri-phones and the states of these tri-phones (senones) are modeled using GMMs [10]. Despite the advantageous properties of GMMs such as faster convergence and capability to model any probability distribution, GMMs fail to model data on non-linear manifold [10]. Though hybrid acoustic models i.e., HMM-Deep neural network (HMM-DNN) have performed better than HMM-GMM systems, the frame level senone labels required for training DNNs have to be obtained from an HMM-GMM system [11]. The hybrid systems suffer from a downside that the objective function which is optimized while training is much different from the true error measure of ASR system (Sequence level transcription accuracy) [12]. Advancements in deep neural networks have greatly influenced the performances of speech recognition systems. Recent developments such as connectionist temporal classification objective function and attention mechanism have enabled end-to-end training for developing acoustic models [12, 13, 14]. End-to-end networks have enriched acoustic models to train without any pre-trained alignments between the acoustic sequence and the label sequence. End-to-End training reduces the mismatch between the true error measure of the system and the objective function which is optimized while training. Apart from the theoretical advantages, end-to-end networks require large amounts of data to train and generalize well. Studies have shown that in the presence of larger sized datasets the performance of end-to-end systems is equivalent to hybrid systems using a pronunciation model and language model [15, 16, 17]. Recent Subspace mixture model has performed superior to traditional speech recognition systems, they have exhibited efficient parameter estimation in limited data scenarios [18, 19].

Multilingual ASR using global phone-sets have been studied in [20, 21, 22]. Sharing some acoustic model parameters have been explored for training multilingual speech recognition [23, 24, 25]. Multi-task architectures have been explored for training a multilingual speech recognition, a hat swap architecture has been mostly explored where lower layers are shared across languages but the higher layers are specific to a language [24, 25]. Subspace Gaussian mixture models have been explored for multi-lingual speech recognition by sharing the subspace defining parameters shared across the languages [18]. The efficiency of grapheme, phoneme-based multilingual speech recognition systems have been studied using RNN-CTC based acoustic models in [26], language feature vectors have been employed in addition to features to condition

10.21437/Interspeech.2018-1584

the systems on language identity. An adaptation mechanism by learning the hidden unit contribution have been explored for multi-lingual and cross-lingual adaptation methods [27]. Recently multilingual ASR for 9 Indian languages comprising 1500 hrs of data has been presented in [28] using Listen attend and spell (LAS) architecture. This model uses a union of monolingual phone-sets comprising of 960 characters to train a single unified model by jointly optimizing acoustic model, pronunciation model, and language model. In-spite of using the union of phone-sets the joint-model has performed better than monolingual models due to the availability of large data for optimizing the model. Most of the studies that have explored multi-task architectures have used multi-lingual data to train monolingual systems or systems with certain parameters shared across languages. In an operating environment, either a front-end language identification (LID) system has to be used to switch to the corresponding monolingual-acoustic model or the best possible hypothesis from all the monolingual models have to be chosen. The former approach demands a front end LID to be very accurate and robust, and the latter requires all the monolingual systems to be operated in parallel. Operating these systems in code-mixed environments gets really complex and challenging. An acoustic model that can seamlessly handle multiple Indian languages without any prior information of the language is required. This study considers the use of a common phone-set as an efficient approach for handling multiple languages in a single system. This study explores acoustic modeling approaches that are more suitable to train a joint acoustic model for Indian languages using common phone-set.

The remaining paper is organized as follows: Section 2 describes the database and the speech recognition frameworks used in this study. Experiments, results, and discussion are presented in section 3. Conclusion and future scope are presented in section 5 and section 6.

2. Database & Experimental setup for end-to-end speech recognition system

2.1. Database

The database is provided by Speech Ocean.com and Microsoft which is released as a part of "Low Resource Speech Recognition Challenge for Indian languages-Interspeech 2018". The database comprises of data from three different languages i.e., Telugu Tamil and Gujarati. The dataset comprises of a 40 hour training set and 5 hour testing set.

2.2. Common phone-set

In this study, a common phone-set was used which is a shared representation across languages. A parser to convert utf8 to IT3 [29] has been used to convert the text to the IT3-format [7]. The text in IT3 is used to generate the pronunciation sequences for all the words. All the multilingual ASR systems used in this study are trained using common phone-set.

2.3. Experimental setup

In this work, an end-to-end ASR has been developed using deep bidirectional Long short term memory networks (LSTMs) [30] using connectionist temporal classification (CTC) [11] objective function.

2.3.1. Deep-bidirectional LSTMs

For the input sequence $S = (s_1, s_2, s_3, ..., s_T)$ the sequence of hidden states computed by an bidirectional-LSTMs layer is given by $H = (h_1, h_2, h_3, ..., h_T)$. At each time step t the forward and backward hidden outputs are concatenated and used as input to the next layer i.e. $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The hidden state sequence is computed using the following equations.

$$i_t = \sigma(W_{is}s_t + W_{ih}h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{fs}s_t + W_{fh}h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_{os}s_t + W_{oh}h_{t-1} + b_o) \tag{3}$$

$$\tilde{c_t} = \tanh(W_{cs}s_t + W_{ch}h_{t-1} + b_c) \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t} \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

The activations of input gate, forget gate, output gate and memory memory cells are given by i_t, f_t, o_t, c_t respectively. The weight matrices $W_{.s}$ connect inputs with the units, where as $W_{.h}$ connects the previous hidden states with the units.

2.3.2. CTC-objective function

Connectionist temporal classification (CTC) is an objective function used to align two sequences of different length [11]. An end-to-end speech recognition system can be trained using CTC objective function and it would not require any frame level alignment between the acoustic sequence and the label sequence. An additional $blank$ label is added to the set of target labels and the probability of not emitting any label at a particular time step is represented using a $blank$ label. As the acoustic sequence and the label sequence are of different lengths an intermediate representation called CTC-$path$ is used to learn the alignment between the acoustic sequence and label sequence. CTC-$path$ gives the target label sequence at frame level and is obtained by repeating the non-blank labels, inserting a blank between two different non-blank labels. The target label sequence is represented by the set of all possible CTC-$paths$.

An input sequence of $X = (x_1, x_2, x_3, ..., x_T)$, the probability of a label sequence being L is obtained by summing the conditional probability $P(l|X)$ over all possible CTC-$paths$.

$$P(L|X) = \sum_{\hat{l} \in \Omega(L)} P(\hat{l}|X) = \sum_{\hat{l} \in \Omega(L)} \prod_{t=1}^{T} P(\hat{l_t}|x_t) \tag{7}$$

Here $\Omega(y)$ is the set of all possible CTC-$paths$. The conditional probability of a label at each time step, $P(\hat{l_t}|x_t)$ is estimated using the network. The network is trained using gradient descent to maximize equation 7, and the forward-backward algorithm is employed to compute gradients [11].

2.4. Subspace-Gaussian mixture models

In a conventional acoustic model, states of HMM are modeled using GMMs. A high-dimensional super-vector of GMM parameters from all the states is expected to lie on a low dimensional manifold common to all the states [19]. Though SGMM uses the GMM as its underlying distribution, but the parameters in an SGMM are shared across the states. These parameters describe the sub-space of the GMM parameters. The individual states can be described using relatively low-dimensional vectors which are the coordinates in the subspace. SGMM can be seen as a compact representation for HMM state distributions. SGMMs perform significantly better than HMM-GMM.

In limited data scenarios, SGMMs have delivered much better performances [18].

3. Experiments, Results & Discussion

Two different acoustic modeling approaches have been explored for training a joint acoustic model i.e, SGMM and RNN-CTC. Recipes from Kaldi-toolkit have been used for training SGMM models. The SGMM-models trained during the study have used 8000 sub-states, and a diagonal UBM of 400 dimensions. End-to-end speech recognition system in this study, has employed deep bidirectional long short-term memory networks (Bi-LSTMs) optimized using CTC objective function. A hyperparameter search has been performed to obtain optimal architectural choices. It has been observed that Deep Bi-LSTMS layers with 3, 4 hidden layers are optimal for training monolingual and joint acoustic models respectively, each layer comprised of 320 units. A learning rate of 0.0001 is used with a batch size of 1. The learning rate is reduced by a factor of 0.5 when a decrease in the validation accuracy is observed. RNN-CTC networks are optimized using Adam optimizer [31] with exponential decays on first and second order momentums are given by 0.9 and 0.99 respectively. The performances of various acoustic models are presented in Table. 1. Monolingual acoustic models trained using SGMM model has been presented in row 2. Row 3 is performance obtained using the joint acoustic model trained using the data from all the three languages. Performances of monolingual acoustic models trained using RNN-CTC based acoustic models have been presented in row 4.

Table 1: *Performances of speech recognition systems trained during the study.*

	Dev set			Eval set		
Acoustic model	Telugu	Tamil	Gujarati	Telugu	Tamil	Gujarati
Monolingual-SGMM	21.69	19.63	14.51	21.75	19.36	21.73
Joint-SGMM	26.53	24.65	17.41	26.22	24.77	26.14
Monolingual-CTC	21.68	21.10	15.07	21.80	20.90	22.94
Joint-CTC	21.28	21.12	14.86	21.73	20.73	21.98
Joint-CTC- Residual connections	21.25	21.07	14.80	21.69	20.87	21.93
Joint-CTC- Gaussian noise	21.25	21.10	14.82	21.49	20.66	21.97
Joint-CTC- Language ID	21.41	20.54	14.62	21.34	20.63	21.64
Joint-CTC- Sub-samp+Gaussian noise	21.26	20.53	14.40	21.52	20.44	21.62
Joint-CTC- Sub-samp+Gaussian noise mono-LM	**20.61**	**20.16**	**14.19**	**20.55**	**19.90**	**21.07**
Context-dependent phones +Sub-samp +Gaussian noise	21.11	20.18	14.75	21.32	19.95	21.91
Context-dependent phones +Sub-samp+ Gaussian noise+mono LM	20.65	19.82	14.61	20.71	19.58	21.69

Apart from handling the acoustic variabilities due to the language, a multilingual speech recognition should handle different orthographies of various languages. As Indian languages share same phonetic space, there can be words with same pronunciation in different languages. When different orthographies are used in the system with a common phone set, this word-phone sequence pairs stand as different entities in the pronunciation model. Such words could be erroneously decoded even when the acoustic model has produced the correct phone sequence. This could be efficiently avoided by using text in IT3 format [29]. IT3-format are any other language independent mapping which could map the words in different languages with same phone sequence as a single entity would be more beneficial in training a multilingual ASR. In this work, we have considered IT3 as the language independent phone sequence based representation. In the present database, out of 140K words, there are 2K words with same phone sequences but different orthographies due to different languages. The performance of joint acoustic models trained using RNN-CTC has been presented in row 5 of Table 1. The transcriptions from training utterances in IT3-format have been used to train a trigram language model. The pronunciation model contains unique words from all the three languages in IT3-format and the corresponding phone sequences.

Residual connections in neural network architectures have lead to a better convergence [32, 33]. In this study, a joint acoustic model has been trained using B-LSTMs with skip connections between two successive hidden layers. The use of these skip connections have eased the convergence during the start of the training but the performance gains are less significant. Use of skip connections have increased the training time. The performance of joint acoustic model using residual connections has been presented in row 6 of Table. 1. Due to a huge increase in training time, residual connections have not been used further in the study. For regularizing the network, 10% of the randomly sampled training examples are chosen and white Gaussian noise (σ=0.075) is injected to these features [34]. The performance of these networks is presented in row 7 of Table.1.To condition the joint acoustic model with the language identity, one-hot language representative vector has been used in tandem with the features [28] and this system has reduced the word error rates but not significantly. The performance of joint acoustic model conditioned on language identity is presented in row 8 of Table. 1.

RNN-CTC based acoustic model is trained to align two sequences of different lengths, unlike the conventional models RNN-CTC does not require any alignment from a pre-trained model. RNNs being sequential in nature reducing the sequence length has reduced the training time, this has been achieved by using pyramidal architectures [35, 14, 17]. Sub-sampling the acoustic sequence by a factor of 2 or 3 has not shown any degradation in the performance of RNN-CTC based acoustic models. In this work, features from successive acoustic frames are concatenated reducing the sequence length by a factor of 2 and the performance obtained by the sub-sampling has been presented in row 9. Performance of speech recognition systems using a joint acoustic model, a common pronunciation model and the language model specific to that language is presented in row 10 of Table. 1. In this work, an RNN-CTC joint acoustic model has been trained to model context-dependent phones ie., phones which occur at starting middle and end of the words, singletons are considered as independent tokens. RNN-CTC is trained it to minimize the token error rate where the tokens used are context-dependent phones and the results are presented in row 11, 12 of Table 1. Though the token accuracy of this systems is 7% lesser but this has produced the WER comparable to the joint acoustic model trained with context independent phones. Use of context-dependent phones has helped in pruning out competing decoding paths. But using the context-dependent phones has increased the number of tokens by a factor of 4 and this has lead to an increase in training time i.e., time for computing CTC loss.

In this study, various approaches for sub-sampling the acoustic sequence has been explored and the results are tabulated in Table. 2. To subsample the acoustic sequence by a factor of 2 alternate frames can be dropped or successive frames can be appended, the performances attained by this sub-sampling methods are presented in row 2, 5 of Table.2. Using a frame shift of 20 ms and a frame size of 30 ms for computing the features would also reduce the acoustic sequence by a factor of 2 and the performance obtained by this sub-sampling has been presented in row 3. The training data can be augmented by a factor of 2 dropping even and odd frames alternatively such sampling is termed as Augmented-sub-sampling. Augmented-sub-sampling would reduce the sequence length and also augment the dataset.

Table 2: *Various approaches for sub-sampling the acoustic sequence.*

	Dev set			Eval set		
Sub-sampling	Telugu	Tamil	Gujarati	Telugu	Tamil	Gujarati
Dropping	22.31	21.28	15.12	22.46	21.01	22.64
Frame-shift 20 ms	21.97	21.43	14.94	21.97	21.17	22.73
AUG-sub-sub-sampling	21.90	21.21	14.84	21.84	21.03	22.42
Appending frames	21.45	20.67	14.72	21.42	20.66	21.97

3.1. Results & Discussion

The performance of joint acoustic models trained using HMM-SGMM is poorer than the performance of HMM-SGMM monolingual models. Using HMM-SGMM for training a joint acoustic model has lead to an increase in word error rate (WER). Indian languages share same phonetic space and differ in phonotactics. The tri-phones modeled by HMM do not share common distribution across different languages. This has lead to the poor performance when a joint acoustic model is trained using HMM-SGMM. Performance monolingual RNN-CTC based acoustic models is less than the monolingual HMM-SGMM system which is in accordance with the earlier studies. The joint acoustic model trained using RNN-CTC has performed better than the monolingual systems. Unlike HMMs, RNN-CTC acoustic models are trained to model context independent phones. The variabilities due to multiple languages have been effectively handled using RNN-CTC.

In the earlier studies, it has been observed that conditioning the acoustic model on language identity (language ID) has improved the performance [28, 26]. Mostly the systems have used the global phone-set which is a union of phone-sets from all the languages. When a joint acoustic model is trained to predict the labels from global phone-set the information about the language identity has improved the performance. It has been observed that the systems trained in such fashion more faithful to language ID, upon encountering a wrong language ID the system has transliterated the acoustic sequence using the phone sequence corresponding to the mismatched language ID. The joint acoustic models trained in this study use a phone-set which is same for all the languages, unlike the union of monolingual phone-sets. Conditioning the model to language ID has helped in convergence but has not significantly improved the performance. Sub-sampling the acoustic sequence in and end-to-end ASR has reduced the training time, this reduction of frames has been explored by dropping the alternate frames, using pyramidal architectures where the sequence length gets reduced along the depth of the network. It has been observed that sub-sampling the acoustic sequence by a factor >3 has affected the convergence of the models.

Decoding the test utterance using a monolingual language model with the joint acoustic model has performed better than the system using the combined language model. Rather than conditioning the joint acoustic model on the language identity and using front-end language identification system. The LID systems developed using phonotactics and syntax of a language are more robust and reliable [36]. LID decision derived from phonotactics of the joint acoustic model could be used to select the corresponding language model to decode the test utterance. A multi-pass decoding could also be a viable solution using a common language model initially to obtain an initial hypothesis and which could give information about the language of the hypothesized test utterance. A second pass decoding with a monolingual model or re-scoring the lattices with the monolingual language model could be beneficial in building multilingual ASR systems for Indian languages.

4. Conclusion

A joint acoustic model is an effective solution for training a multi-lingual speech recognition system, rather than using a front-end LID to switch between monolingual models or using parallel monolingual models. In an under-resourced scenario, use of a common phone-set could be an efficient approach for sharing data across the languages. This work studies the amenability of various acoustic models i.e., HMM-SGMM and RNN-CTC for developing a joint acoustic model using common phone-set. It has been observed that end-to-end systems which model context independent phones as a basic unit have performed better than HMM-SGMM system which models context dependent tri-phone. RNN-CTC based acoustic models have shown to be more promising while using common phone-set. Conditioning the joint acoustic model model with language ID has not improved the performance significantly. Converting orthographies of various languages to IT3-format can be helpful in handling the words in different languages with same pronunciation sequences. Use of a joint acoustic model and text in IT3 format could be a viable solution to operate multilingual and code-mixed speech recognition systems irrespective of input languages. Using a monolingual language model in a multi-pass decoding framework would improve the performances significantly.

5. Future scope

Apart from multiple advantages, training end-to-end networks need large sized datasets for better generalization. Recent advancements in neural networks such as zone-out, use of variational Bi-LSTMS could improve the performances of joint acoustic models. Architectures such as LAS, use of multi-Head attentions, minimum WER based training could improve the performances significantly. Apart from performance, the decoders in sequence-to-sequence models are auto-regressive in nature and this would increase the latency of the systems. Recent studies which use the latent variables rather than auto-regression could reduce the latency in decoders.

6. References

[1] Rohit Kumar, S Kishore, Anumanchipalli Gopalakrishna, Rahul Chitturi, Sachin Joshi, Satinder Singh, and R Sitaram, "Development of indian language speech databases for large vocabulary speech recognition systems," in *Proc. Int. Conference on Speech and Computer*, 2005.

[2] A Nayeemulla Khan, Suryakanth V Gangashetty, and S Rajen-

dran, “Speech database for indian languages-a preliminary study,” in *Proc. Int. Conf. Natural Language Processing*, 2002, pp. 295–301.

[3] Gautam Varma Mantena, S Rajendran, B Rambabu, Suryakanth V Gangashetty, B Yegnanarayana, and Kishore Prahallad, “A speech-based conversation system for accessing agriculture commodity prices in indian languages,” in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on*. IEEE, 2011, pp. 153–154.

[4] Aanchan Mohan, Richard Rose, Sina Hamidi Ghalehjegh, and S Umesh, “Acoustic modelling for speech recognition in indian languages in an agricultural commodities task domain,” *Speech Communication*, vol. 56, pp. 167–180, 2014.

[5] C Santhosh Kumar, VP Mohandas, and Haizhou Li, “Multilingual speech recognition: A unified approach,” in *Proc. European Conf. Speech Communication and Technology*, 2005.

[6] Kishore Prahallad, E Naresh Kumar, Venkatesh Keri, S Rajendran, and Alan W Black, “The IIIT-H Indic speech databases,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] Arun Baby, NL Nishanthi, Anju Leela Thomas, and Hema A Murthy, “A unified parser for developing indian language text to speech synthesizers,” in *International Conference on Text, Speech, and Dialogue*. Springer, 2016, pp. 514–521.

[8] B Ramani, S Lilly Christina, G Anushiya Rachel, V Sherlin Solomi, Mahesh Kumar Nandwana, Anusha Prakash, S Aswin Shanmugam, Raghava Krishnan, S Kishore Prahalad, K Samudravijaya, et al., “A common attribute based unified hts framework for speech synthesis in indian languages,” in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[9] Ayushi Pandey, Brij Mohan Lai Srivastava, and Suryakanth V Gangashetty, “Adapting monolingual resources for code-mixed hindi-english speech recognition,” in *Proc. IEEE International Conference on Asian Language Processing*, 2017, pp. 218–221.

[10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., “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.

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proc. International Conference on Machine Learning*. ACM, 2006, pp. 369–376.

[12] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, “Attention-based models for speech recognition,” in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[13] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, “End-to-end attention-based large vocabulary speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2016, pp. 4945–4949.

[14] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, “Listen, attend and spell,” *arXiv preprint arXiv:1508.01211*, 2015.

[15] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, “Fast and accurate recurrent neural network acoustic models for speech recognition,” *arXiv preprint arXiv:1507.06947*, 2015.

[16] Haşim Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk, “Learning acoustic frame labeling for speech recognition with recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*. IEEE, 2015, pp. 4280–4284.

[17] Golan Pundak and Tara N Sainath, “Lower frame rate neural network acoustic models.,” in *Proc. INTERSPEECH*, 2016, pp. 22–26.

[18] Lukáš Burget, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Daniel Povey, et al., “Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*. IEEE, 2010, pp. 4334–4337.

[19] Daniel Povey, “A tutorial-style introduction to subspace gaussian mixture models for speech recognition,” *Microsoft Research, Redmond, WA*, 2009.

[20] Tanja Schultz and Alex Waibel, “Fast bootstrapping of lvcsr systems with multilingual phoneme sets,” in *Proc. European Conf. Speech Communication and Technology*, 1997.

[21] Tanja Schultz and Alex Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition,” *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

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

[23] Georg Heigold, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, MarcAurelio Ranzato, Matthieu Devin, and Jeffrey Dean, “Multilingual acoustic models using distributed deep neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*. IEEE, 2013, pp. 8619–8623.

[24] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, “Multilingual training of deep neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*. IEEE, 2013, pp. 7319–7323.

[25] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*. IEEE, 2013, pp. 7304–7308.

[26] Markus Müller, Sebastian Stüker, and Alex Waibel, “Phonemic and graphemic multilingual ctc based speech recognition,” *arXiv preprint arXiv:1711.04564*, 2017.

[27] Sibo Tong, Philip N Garner, and Hervé Bourlard, “Multilingual training and cross-lingual adaptation on ctc-based acoustic model,” *arXiv preprint arXiv:1711.10025*, 2017.

[28] Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao, “Multilingual speech recognition with a single end-to-end model,” *arXiv preprint arXiv:1711.01694*, 2017.

[29] Ganapathiraju Madhavi, Balakrishnan Mini, N Balakrishnan, and Raj Reddy, “Om: One tool for many (indian) languages,” *Journal of Zhejiang University-SCIENCE A*, vol. 6, no. 11, pp. 1348–1353, 2005.

[30] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proc. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] Hari Krishna Vydana and Anil Kumar Vuppala, “Residual neural networks for speech recognition,” in *Proc. IEEE European Signal Processing Conference*, 2017, pp. 543–547.

[34] Alan Graves, Abdel Rahman Mohamed, and Geoffrey Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2013, pp. 6645–6649.

[35] Liang Lu, Xingxing Zhang, Kyunghyun Cho, and Steve Renals, “A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition.,” in *Proc. INTERSPEECH*, 2015, pp. 3249–3253.

[36] Brij Mohan Lal Srivastava, Hari Vydana, Anil Kumar Vuppala, and Manish Shrivastava, “Significance of neural phonotactic models for large-scale spoken language identification,” in *Int. Joint Conf. Neural Networks*, 2017, pp. 2144–2151.