Interspeech 2018
2-6 September 2018, Hyderabad

TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages

Noor Fathima, Tanvina Patel, Mahima C and Anuroop Iyengar

Cogknit Semantics, Bangalore, Karnataka, India
{noorfathima, tanvina, mahima, anuroop}@cogknit.com

Abstract

India is a diverse and multilingual country. It has vast linguistic variations, spoken across its billion plus population. Lack of resources in terms of transcribed speech data, phonetic pronunciation dictionary or lexicon, and text collection has hindered the development and improvement of the ASR systems for Indic languages. With the Interspeech 2018 Special Session: Low Resource Speech Recognition Challenge for Indian Languages, efforts have been made to solve this issue to an extent. In this paper, we explore the fact that the shared phonetic properties of the languages are essential for improved ASR performance. We build a multilingual Time Delay Neural Network (TDNN) system that uses combined acoustic modeling and language-specific information to decode the input test sequences. Using this approach, for Tamil, Telugu and Gujarati language we obtain a Word Error Rate (WER) of 16.07%, 17.14%, 17.69%, respectively, which was the second best system at the challenge.

Index Terms: Speech recognition, low-resource languages, multilingual systems, TDNN

1. Introduction

According to Census of India of 2001, India has 22 major languages and 1599 other languages. Out of these, 30 languages are spoken by more than a million people and 22 languages have been accorded with the ”official language” status [1]. With the growing use of the Internet and with the idea of digitization taking center stage, speech technology applications in Indian languages will play a crucial role in the development of various domains like agriculture, health care and availing government services by common men, etc. [2]. Voice User Interface (VUI) devi are finding their way into everyday life all over the world. In a country like India where a sizeable population is unlettered and hence, often technologically novice, developing resource robust spoken language technologies has the potential to make a difference between a good quality of life and a bad one. Therefore, developing speech recognition based technologies has tremendous prospects in the country.

In the development of spoken language technologies like Automatic Speech Recognition (ASR) systems, low resource factor poses a serious challenge. Majority of the languages in India are low resourced. In the context of ASR, the term 'low-resource', although not strictly defined, usually refers to the lack of availability of transcribed audio data, lack of extensive linguistic research in the language which in turn is a factor for the lack of a pronunciation dictionary, a key ingredient in ASR systems. Hence, there is an urgent need to mitigate the problem. Traditional monolingual ASR systems require thousands of hours of transcribed audio data in order to produce an acceptable performance in terms of Word Error Rate (WER) [3]. For low-resourced languages, such costly requirements can be overcome by considering techniques such as word decompounding which leverages the available lexical data to improve ASR performance [4]. Recently, advances made in Deep Neural Network (DNN) architectures and multilingual training techniques have outperformed and continue to show promising results.

In this paper, we develop ASR systems for low-resourced languages as a part of the 'Interspeech 2018: Low Resource Speech Recognition Challenge for Indian Languages' [5]. The aim is to develop ASR systems for three languages: Gujarati, Tamil and Telugu. Of these languages, Gujarati is an Indo-Aryan language while Tamil and Telugu are Dravidian languages. Both monolingual and multilingual systems were developed. It was observed that multilingual systems provided better performance. This keeps us on the similar lines with the research on development of ASR systems for low resource languages [6, 7]. The rest of the paper is organized as follows, Section 2 describes the basic speech recognition system, Section 3 talks about the multilingual systems and the approach used in the paper. Section 4 describes the experimental setup and details of the parameter settings and performance measures used. Section 5 describes the experimental results and discussions on it. Finally Section 6 concludes with future research directions.

2. Speech Recognition System

2.1. Language Model (LM)

The Language Model (LM) estimates the probability of a hypothesized word sequence, or LM score, by learning the correlation between words from a training text corpora. The LM score often can be estimated accurately if the prior knowledge about the domain or task is known [8]. Text data corresponding to the training audio-set in its normalized form was provided by the challenge organizers. The SRI Language Modeling (SRILM) toolkit [9] was used to train Kneser-Ney smoothed trigram LMs on the training text data of each language. The Lexicon for each language uses a Common Label Set [10, 11] where phones of multiple languages are mapped to a universal phone-set.

2.2. Acoustic Model (AM)

Based on the available common phone-set we train language independent multilingual Acoustic Models (AMs) [7]. These include Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM), hybrid Deep Neural Network (DNN)-HMM and Time Delay Neural Networks (TDNN).

2.2.1. GMM-HMM systems

In the 1980s, state-of-the-art ASR systems used Mel-Frequency Cepstral Coefficient (MFCC) or Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP) [12, 13] as feature vectors along with GMM-HMM. These GMM-HMM AMs

 10.21437/Interspeech.2018-2117

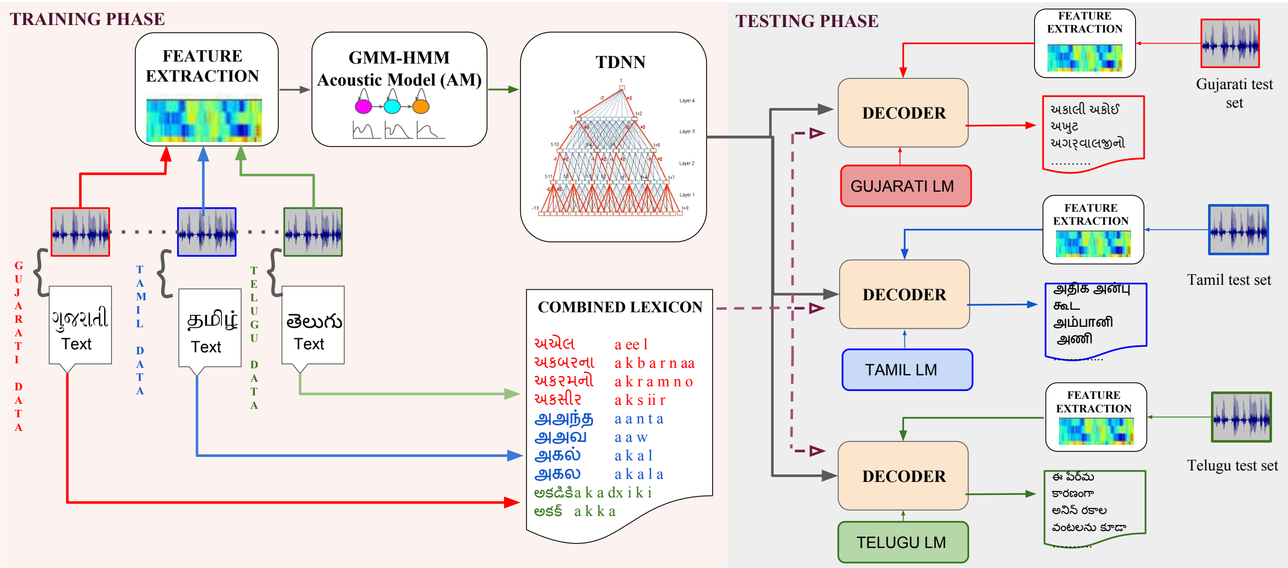

Figure 1: *Architecture of ASR system submitted for the IS 2018 Challenge*

were trained using the Maximum Likelihood (ML) training criterion. Later, in the 2000s, sequence discriminative training algorithms such as Minimum Classification Error (MCE) and Minimum Phone Error (MPE) were proposed that further improved the ASR accuracy [14]. In this work, MFCC features are extracted with the Δ and $\Delta\Delta$ features for initial speaker independent GMM-HMM training. The speaker dependent GMM-HMM model is built using Feature space Maximum Likelihood Linear Regression (FMLLR) features [15].

2.2.2. DNN-HMM systems

Over the last few years, efficient methods for training DNNs for ASR have been witnessed [16] showing that DNNs are better classifiers than GMMs [17]. The output layer accommodates the number of HMM states that arise when each phone is modeled by a number of different triphone HMMs taking into account the phones on either side [18]. The GMM-HMM model is used to obtain alignments for training data and finally a DNN is trained by feeding the FMLLR features as the input and senone probabilities as the output. The DNN training uses p-norm activation function [19] with cross entropy as the loss function using natural Stochastic Gradient Descent (SGD) [20].

2.2.3. TDNN systems

Time Delay Neural Networks (TDNN) have shown to be effective in modelling long-range temporal dependencies [21]. This property of TDNNs is exploited by training AMs that can learn the long term dependencies based on short-term feature representations [22]. Along with 40-dimensional MFCCs, 100-dimensional iVectors are appended at each time step [23]. It has been observed that iVectors capture both speaker and environment specific information and have been shown to be useful for instantaneous and discriminative adaptation of the neural network. Training data is artificially increased by 3-fold through time-warping of raw audio [24]. The training procedure for chain models is a Lattice-Free (LF) version of Maximum Mutual Information (MMI) criterion without the need for frame-level cross-entropy pre-training [25].

3. Multilingual Systems

Before neural networks were used in Large Vocabulary Continuous Speech Recognition (LVCSR) task, techniques such as Multilingual-Mix or Multilingual-Tag have been used to exploit data from multiple languages for GMM-HMM systems [7]. Multilingual DNNs for speech recognition systems have shown to provide consistent advantages for low resource languages. Through multilingual acoustic modeling, the acoustic data is shared across multiple languages. This ensures a greater and hence, a better coverage of contextual variation in all languages being considered. This improved coverage is ensured especially due to the usage of a universal phone-set that maps acoustic units of different languages to language-independent speech units [26]. The hidden layers are hypothesized to be global feature extractors [27]. The languages that are pooled together to form the training data set are usually referred to as source languages and the language for which the system will be tested is referred as the target language. This approach also reduces the time required for training.

The overall architecture and theme of the system that was used to develop the ASR system under the constraints specified in the challenge is as shown in Figure 1. The source languages here are Gujarati, Tamil and Telugu. The training phase uses combined audio data of approximately 120 hours along with corresponding transcripts. Using MFCCs without cepstral truncation, speaker dependent GMM-HMM system is built using FMLLR features. The alignments obtained from this model is used for LF-MMI training of the chain TDNN. A combined lexicon covering words occurring in training set of all three languages is used. During the testing phase, features are extracted from the test audio and language-specific language model is used for decoding to obtain output transcriptions for each language. While this system did give promising results in terms of improved WER, the ASR system built using Tamil and Telugu as source languages gave best performance when decoded individually for Tamil and Telugu using their specific language models. We trained a separate chain TDNN for Gujarati language which gave us better performance.

Table 1: *Results in %WER on the test set for training in one language and decoding using LM of the same language*

	GMM-HMM		DNN-HMM		TDNN	
Training	**Baseline***	**Proposed**	**Baseline***	**Proposed**	**Baseline***	**Proposed**
Gujarati	23.78	16.95	27.79	14.38	19.76	12.7
Tamil	33.55	20.68	25.47	18.04	19.45	16.35
Telugu	40.12	24.24	34.97	21.05	22.61	18.61

* Baseline results are provided by the challenge organizers [5]

4. Experimental Setup

4.1. Corpus Details

The corpus was provided by the organizers of the Low Resource Speech Recognition Challenge for Indian Languages [5]. Training data of 40 hours along with the corresponding text transcriptions and lexicon were provided for each of the three languages. As a rule of the challenge, only the data provided by the challenge should be used to build the systems (although the choice to use the corresponding language's data to build each system or combine the data and use it cross-lingually was provided). The evaluation was carried out using an additional 5 hours of test data provided (as shown in Table 2). The test data can be used for parameter tuning etc. The final test known as the blind test was provided for which the results were provided by the challenge organizers on submission of the hypothesis.

Table 2: *Details of the training and testing data [5]*

	Training		Test	Blind Test
Languages	**# of Utt.**	**Lexicon (Words)**	**# of Utt.**	**# of Utt.**
Gujarati	22807	43576	3075	3419
Tamil	39131	57745	3081	2609
Telugu	44882	48680	3040	2549

4.2. Parameter Settings

We use 3 hidden layer DNN for model training. Each hidden layer has 2000 dimensional hidden units with p-norm activation. Here, p=2 and group size = 5 which leads to input p-norm dimension 2000 and output p-norm dimension 400. The DNN has input layer which takes 360-dimensional input and the output of the DNN is 2365 context dependent phonemes states. The input to the neural network is obtained by concatenating 4 left and 4 right FMLLR features each of dimension 40. The outputs are obtained by GMM-HMM alignment. We minimize the cross-entropy loss function using back-propagation with an initial learning rate of 0.01 and final learning rate of 0.001.

The chain TDNN model was composed of 6 layers with 725 Rectified Linear Units (ReLU) at the input layer. The input features are at the original frame rate of 100 per second and the output frame rate is reduced by 3-fold. The first splicing is removed before the Linear Discriminant Analysis (LDA) transforms layer. The spliced indices in the consecutive layers were [-1,0,1,2;-3,0,3;-3,0,3;-3,0,3;-6,-3,0] with LDA applied to the input features. The training was carried out using GeForce GTX 1080 Ti (with 11GB RAM and Cuda cores = 3584).

4.3. Tools and Performance Measures

The Kaldi Speech recognition toolkit is used [28]. The LibriSpeech recipe was used for all experiments including acoustic feature extraction, training and decoding [29]. The SRILM toolkit was used for language modeling [9]. Modified Kneser-Ney smoothed 3-gram models were built having the least perplexity. The performance of ASR is estimated using the Word Error Rate (WER) as the measure (as decided for the challenge).

5. Experimental Results

In this section, we describe in detail the experiments conducted and the submission made towards the challenge [5]. The experiments are carried for GMM-HMM, DNN-HMM and TDNN based systems. The results and the comparative study with the baseline results are shown in the next sub-sections.

5.1. Training with One Language

In this section we show the results of the experiments conducted as per the baseline of the challenge [5]. The challenge baseline uses TIMIT recipe from the Kaldi toolkit to report the GMM-HMM and DNN-HMM results and Wall Street Journal (WSJ) recipe to report the TDNN results. In our approach, we use the LibriSpeech recipe for all the three architectures. The results are shown in Table 1. It is observed that for all the languages and for all the systems, the WERs obtained where less than that of the baseline. For Gujarati, the WER is decreasing gradually from GMM (16.95 %), DNN (14.38 %) to TDNN (12.7 %) systems. However, for the baseline systems, the WER increases for DNN as compared to that of the GMM-HMM system. A similar decreasing trend of WER is also observed for Tamil and Telugu. The WER using TDNN architecture was very low for Gujarati and around 16-18% for Tamil and Telugu.

5.2. Training with Multiple Languages

The focus of the present work is to highlight the fact that Indian languages are similar to each other and using acoustic data from multiple similar languages should improve the performance of the ASR system. Therefore, we first combine the acoustic data and next experiment with the lexical content for improved performance as described next.

5.2.1. 3-Language Training (3-lang)

In this method, we train the system by combining the acoustic training data from each of the three languages. To understand the effect of the LM, we merge the lexicons and text of the 3 languages and build a mixed LM. On decoding, we decode the test language using this mixed LM. The results for the same are shown in Table 3. It is observed that the performance of Gujarati improved from 12.7% (for TDNN as in Table 1) to 11.95%. On the other hand, for Tamil and Telugu, the WER seemed to degrade more for GMM-HMM and DNN-HMM systems as compared to TDNN systems trained on a single language.

It is known that although the languages may have similar sounding phones, the lexical content is different. The issue of decoding with mixed LM is that the ASR output is likely to have the text of all the languages mixed. Hence, it may not be readable as there may be other scripts as compared to that of the target language. Also, the lexical mixing should be carried if the speech data has many languages spoken. Therefore, for the next set of experiments, we decode the test data using the LM of the target language only (i.e., language-specific decoding). Table 3 shows that on decoding with the language-specific LMs (and using the same 3-lang AM), the WER decreases as compared to using mixed LM. The decrease is around 1% for all the architectures and almost for all languages.

Table 3: *Results in %WER on training with 3 languages and decoding with mixed LM as well as language specific LM*

LM	Lang.	GMM-HMM	DNN-HMM	TDNN
Mixed	**Guj.**	18.87	14.18	11.95
	Tamil	24.81	19.25	17.32
	Telugu	27.19	21.21	18.84
Guj.	**Guj.**	17.28	13.43	11.58
Tamil	**Tamil**	22.46	18.5	16.79
Telugu	**Telugu**	24.21	19.85	17.94

5.2.2. 2-Language Training (2-lang)

It is known that Tamil and Telugu belong to the same Dravidian language family as compared to that of Gujarati which belongs to the Indo-Aryan family. Also, it is evident from the Table 3 that the WERs of Tamil and Telugu (on combining the acoustic information of 3-languages) did not decrease significantly than that of the individual systems as in Table 1. Thus, we conduct the same experiment of combining the acoustic information, however, this time only for languages that fall under the same language family (i.e., Tamil and Telugu).

Table 4: *Results in %WER on training with Tamil and Telugu and decoding with mixed LM as well as lang. specific LM*

LM	Lang.	GMM-HMM	DNN-HMM	TDNN
Mixed	**Tamil**	22.94	19.27	16.46
	Telugu	24.47	20.28	17.69
Tamil	**Tamil**	24.03	19.1	16.47
Telugu	**Telugu**	25.84	20.36	16.95

As shown in Table 4, for TDNN systems and on using mixed decoding, the WER decreases for Tamil language to 16.46 % which was 17.32 % when trained on 3-lang AM. A similar observation was found for Telugu. On decoding with language-specific LM, the WER decreases from 16.79% to 16.47% for Tamil and 17.94% to 16.95% for Telugu. This shows that combining similar acoustic data for training and decoding with the LM of the target language results in improved performance.

5.3. Best Results on the Blind Test

For the Interspeech 2018 challenge, we submit 6 hypothesis (2 for each language) on the blind test set as follows:

Hypothesis 1

- Gujarati: TDNN with Gujarati AM and Gujarati LM
- Tamil: TDNN with Tamil AM and Tamil LM
- Telugu: TDNN with Telugu AM and Telugu LM

Hypothesis 2

- Gujarati: DNN with 3-lang AM and Gujarati LM[1]
- Tamil: TDNN with 2-lang AM and Tamil LM
- Telugu: TDNN with 2-lang AM and Telugu LM

The results for the hypothesis on the blind test set are shown in Table 5. The first hypothesis for Gujarati and second hypothesis for Tamil and Telugu were ranked the second best system for the challenge. These results are shown in bold in Table 5. On decoding with the mixed LM, it is observed that the % WER is very high for all the languages. This was not observed in Table 3 where the WER on decoding with mixed LM and language-specific LM were almost same. Thus, our choice of the submissions for the challenge were made correctly. Thus, using multilingual acoustic data and decoding with language-specific LM seems to generalize for the given test data.

Table 5: *Results in %WER on the blind test data*

AM	LM	Language	TDNN
Gujarati	Gujarati	Gujarati	**17.69**
Tamil	Tamil	Tamil	26.99
Telugu	Telugu	Telugu	28.75
Tamil+Telugu (2-lang)	Mixed (2-lang)	Tamil	16.65
		Telugu	17.88
	Tamil	Tamil	**16.07**
	Telugu	Telugu	**17.14**
Gujarati+Tamil +Telugu (3-lang)	Mixed (3-lang)	Gujarati	29.57
		Tamil	20.21
		Telugu	23.72
	Gujarati	Gujarati	16.69
	Tamil	Tamil	16.36
	Telugu	Telugu	17.04

6. Summary and Conclusions

In this paper, we describe our approach used for developing ASR systems as a part of the Interspeech 2018: Low Resource Speech Recognition Challenge for Indian Languages. We demonstrate the significance of using combined acoustic modeling between languages that are similar acoustically. The given test data was not used for training as the blind test was unknown and could have resulted in over-fitting. We conclude that language specific decoding when combined with more acoustic data resulted in improved performance. Using this approach for all the languages we obtain a WER of 16-17%, which is almost uniform for all languages. Hence, all languages performed equally well. Our results also conform with the idea that a source language close to the target language is more beneficial than a random one [3]. As a part of future work, we are interested in improving ASR performance by combining data from Indian languages which are phonetically similar to each other. We also wish to explore and develop a keyword search system for these Indian languages.

7. Acknowledgements

The authors thank the organizers of the special session for organizing the challenge and provide the database for research. We also thank Cogknit for helping us carry out the research work.

[1]TDNN systems were under experimentation during the challenge submission

8. References

[1] Wikipedia. (2018) Languages of India. [Online]. Available: https://en.wikipedia.org/wiki/Languages_of_India

[2] A. Mohan, R. Rose, S. H. Ghalehjegh, and S. Umesh, "Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain," *Speech Communication*, no. 56, pp. 167–180, Jan. 2014.

[3] E. Chuangsuwanich, "Multilingual techniques for low resource automatic speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology. Department of Electrical Engineering and Computer Science, 2012.

[4] T. Pellegrini and L. Lamel, "Investigating automatic decomposition for ASR in less represented languages," in *INTERSPEECH*, Pittsburgh, Pennsylvania, USA, 2006, pp. 285,288.

[5] INTERSPEECH-2018. Special Session: Low resource speech recognition challenge for Indian languages. [Online]. Available: https://www.microsoft.com/en-us/research/event/interspeech-2018-special-session-low-resource-speech-recognition-challenge-indian-languages/

[6] Z. Wang, U. Topkara, T. Schultz, and A. Waibel, "Towards universal speech recognition," in *IEEE International Conference on Multimodal Interfaces (ICMI)*, Washington, DC, USA, 2002, pp. 247–252.

[7] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *International Conference Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998, pp. 1819–1822.

[8] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, ser. Signals and Communication Technology. Springer, 2015.

[9] A. Stolcke, "SRILM-An extensible language modeling toolkit," in *International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, 2002, pp. 901–904.

[10] B. Ramani, L. Christina, A. G. Rachel, S. V. Solomi, M. K. Nandwana, A. Prakash, A. Shanmugam, R. Krishnan, K. Prahalad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *ISCA Speech Synthesis Workshop (SSW)*, Barcelona, Spain, pp. 291–296.

[11] Indian Language TTS Consortium and ASR Consortium. (2016) Indian Language Speech sound Label set (ILSL12). [Online]. Available: https://www.iitm.ac.in/donlab/tts/downloads/cls/cls_v2.1.6.pdf

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[13] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Washington, DC, 1992, pp. 121–124.

[14] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *INTERSPEECH*, Lyon, France, 2013, pp. 2345–2349.

[15] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language (CSL)*, vol. 12, no. 2, pp. 75–98, 1998.

[16] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[17] A. rahman Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, pp. 1–9, 2009.

[18] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[19] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 215–219.

[20] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," 2014. [Online]. Available: http://arxiv.org/abs/1410.7455

[21] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Readings in speech recognition," A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, ch. Phoneme Recognition Using Time-delay Neural Networks, pp. 393–404.

[22] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 3214–3218.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[24] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 3586–3589.

[25] 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*, San Francisco, US, 2016, pp. 2751–2755.

[26] H. Lin, L. Deng, C.-H. Lee, D. Yu, A. Acero, and Y. Gong, "A study on multilingual acoustic modeling for large vocabulary asr," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009, pp. 4333–4336.

[27] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in *INTERSPEECH*, 2017, pp. 714–718.

[28] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, US, 2011, pp. 1–4.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, pp. 5206–5210.