PSEUDO LIKELIHOOD CORRECTION TECHNIQUE FOR LOW RESOURCE ACCENTED ASR

Avni Rajpal, Achuth Rao MV, Chiranjeevi Yarra, Ritu Aggarwal, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

ABSTRACT

With the availability of large data, ASRs perform well on native English but poorly for non-native English data. Training non-native ASRs or adapting a native English ASR is often limited by the availability of data, particularly for low resource scenarios. A typical HMM-DNN based ASR decoding requires pseudo-likelihood of states given an acoustic observation, which changes significantly from native to non-native speech due to accent variation. In order to improve the performance of a native English ASR on non-native English data, we, in this work, propose a DNN-based pseudo-likelihood correction (PLC) technique, in which a non-native pseudo-likelihood vector is mapped to match its native counterpart. Instead of correcting all elements of a non-native pseudo-likelihood vector, a loss function is proposed to correct only top few of them. Experiments with one native and multiple Indian English corpora show an improvement of WER by ~12% and ~5% using the proposed PLC technique over unadapted and adapted native English ASR respectively, when recognition is performed on an Indian English corpus different from that used for both PLC and adaptation. Experiments with upto 2 hours of parallel native and non-native English data reveal that, PLC performs better than adaptation for all unseen cases considered.

***Index Terms*—** Adaptation, Pseudo-likelihood Correction Technique, LF-MMI

1. INTRODUCTION

With the advent of deep learning, automatic speech recognition (ASR) technology has received a boost in its performance. Presently, state-of-the-art ASR systems, based on the dataset and benchmark test sets, are known to achieve 90 - 95 % word level accuracy [1]. However, performance of these ASR systems drastically degrades for non-native users, thereby limiting the use of the technology. Studies have shown that even the most advanced ASR systems show two times more word error rate (WER) for non-native speakers compared to that for native speakers [2, 3, 4]. If we assume the read speech scenario, then the factors that majorly contribute to poor performance of ASR could be 1) absence of pronunciation variations in the lexicon 2) high confusion in the posteriors obtained from the native acoustic model due to unseen accent variations. In this work, we mainly focus on acoustic modelling approaches to improve the ASR performance for non-native speakers. One of the most common approaches for acoustic modelling is to adapt acoustic model through techniques including MAP, MLLR [5, 6, 7]. In [8, 9, 10], a DNN based acoustic model adaptation is used in which hidden layers are shared and accent specific output layers are trained to learn the parameters of the model for non-native speech. When a small amount of data is available for adaptation, then variability for only limited number of senones is observed. As a result, adaptation techniques tend to overfit the data distribution and thus cannot generalize well. Experiments by Ghahremani et al. [11], showed that weight transfer is effective for small amount of target corpus. Moreover, for tiny target corpus as small as 5hours as in MGB-3 challenge, transferring weights of the whole network including final layer was found very useful [9]. Furthermore, to prevent over-fitting, regularization techniques such as drop out or adding additional loss terms such as KL divergence between the outputs of the original and adapted model were introduced [12]. All the above adaptation approaches require transcription to provide ground truth labels. Recently, the teacher/student (T/S) based adaptation was proposed which utilizes the parallel data instead of the transcription [13, 14]. In this approach, posterior probabilities or soft labels generated by the teacher (source) model are used instead of the labels from the transcription to train a student (target) model with the parallel data from the target domain. This approach is found to be useful in the scenarios where huge amount of parallel data is available.

Typically for the HMM-DNN based ASR models, the conventional approaches for adaptation, as described above, modify the source model parameters such that the output distribution becomes close to the target's ground truth distribution. As a result, all the state posterior values need to be corrected and are given equal weights. However, in the ASR decoding process, only top few state score values per frame contribute in obtaining the optimal hypothesis [15]. This suggests that top few state score values should be more accurate than the remaining values, and, hence, the objective function used for the adaptation, should consider this information. In this work, we propose to optimize only top L values for adaptation. Our approach is similar to the T/S based adaptation using parallel data, except that instead of using KLD as the objective function, we use mean squared error (MSE). This is because the output of the Lattice-free Maximum Mutual Information (LF-MMI) based source neural network model, used for decoding, is interpreted as pseudo likelihood [16] and not, the state posterior probability. Hence in this paper, we propose DNN based pseudo-likelihood correction (PLC) mapping that is trained to correct top L values of the non-native pseudo-likelihoods to be as close as the native pseudo-likelihoods. We experimented with two objective functions: 1) one that minimizes MSE defined using top few pseudo-likelihood values from both input (non-native) and output (native) and 2) another that minimizes MSE defined using top few pseudo-likelihood values from the output only.

Two hours of parallel data from native and non-native speakers is used for training the proposed PLC scheme. In addition, 3 different unseen test sets with different recording conditions are used to investigate the robustness of the proposed approach. Experiments reveal that the proposed PLC approach yields significant improvement in word error rate (WER) compared to an unadapted native ASR system. Experiments on unseen test set indicate the robustness of PLC to different recording conditions.

Authors thank the Department of Science and Technology, Govt. of India for their support in this work.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

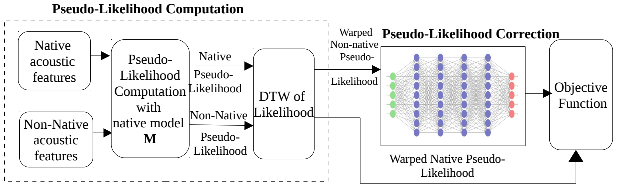
Pseudo-Likelihood Computation
Native acoustic features
Non-Native acoustic features
Pseudo-Likelihood Computation with native model M
Native Pseudo-Likelihood
Non-Native Pseudo-Likelihood
DTW of Likelihood
Warped Non-native Pseudo-Likelihood
Pseudo-Likelihood Correction
Objective Function
Warped Native Pseudo-Likelihood

Fig. 1. Block diagram summarizing the steps of the proposed pseudo-likelihood correction (PLC) approach.

2. DATABASE

In our experiments, Librispeech dataset is used for training base native model [17]. In order to learn PLC, parallel data set is created using native speakers' utterances from TIMIT dataset and Indian English utterances from Indian-TIMIT (iTIMIT) dataset [18]. For robustness related experiments, Indian English speakers' recordings from Voxforge (VOX) [19], Common Voice (MOZ) [20] and Indic Mobile (iMob) dataset are used. All the three datasets iMob, VOX and MOZ are collected through crowd sourcing. Hence, the recordings in these datasets have variable background noise as opposed to the recordings from TIMIT and iTIMIT which are collected in less noisy lab environment. The details of iTIMIT and iMob are described below.

Indic TIMIT (iTIMIT) Indic TIMIT is a database of spoken utterances in English by Indian speakers from different native language backgrounds. The data is collected in our laboratory from a total of 80 subjects, each providing recordings of all 2342 unique sentences from the TIMIT corpus. The speakers in this collected corpus are chosen to have L1 such that it is spoken by majority of the population. More details regarding distribution of the speakers per language as well as per geographical region is provided in [21]. For the recording, a total of 16 subjects from each of the 5 groups as described in [21] with equal males and females are considered in order to maintain uniformity across the groups. The speech samples were collected at a sampling rate of 48kHz with 16bit PCM format, in clean read speech condition. The recordings were later downsampled to 16kHz, to be used for our experiments.

Indic Mobile (iMob) Indic Mobile dataset was collected by us through mobile application with a support of an industry partner. The corpus consists of 100 hours of read speech from 827 Indian speakers from several regions in India. The speakers were asked to read set of sentences that appear on their application in various environments including home, school and market. The prompts were selected to be phonetically balanced. In addition, certain prompts were forced to have Nouns in order to capture pronunciation diversities from a large group of Indian speakers. The audio recordings are dual channel, collected at a sampling rate of 16kHz with 16-bit PCM format.

3. PROPOSED METHOD

Given a sequence of acoustic feature vectors $\mathbf{O}$, the corresponding sequence of words $\mathbf{W}$, in an ASR, is obtained using following equation[15]:

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmax}} \log P(\mathbf{W} \mid \mathbf{O}). \tag{1}$$

It can be shown that the eq(1) can be reformulated as:

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmax}} \log \frac{P_\theta(\mathbf{O} \mid \mathbf{Q}) P(\mathbf{W})}{\sum_{\mathbf{W}'} P_\theta(\mathbf{O} \mid \mathbf{Q}') P(\mathbf{W}')}. \tag{2}$$

where $\mathbf{Q}$ is the sequence of states and θ are the set of parameters of the HMM-DNN model, estimated using Maximum Mutual Information (MMI) criterion[16]. Given a set of training utterances $\{\mathbf{O}_u, \mathbf{W}_u\}_{u=1}^{U}$ and $\mathbf{Q}_u$ being the state sequence corresponding to $\mathbf{W}_u$, it is shown that, for both LF-MMI and conventional MMI training, $P_\theta(\mathbf{O}_u \mid \mathbf{Q}_u)$ is a function of the final layer output of DNN $y(u,t)$ at time frame t and utterance u [22, 23]. Specifically for LF-MMI as shown in [16, 23], the DNN output is interpreted as pseudo-likelihood, therefore, $y(u,t)$ is the DNN output without softmax activation and is directly used as acoustic score during decoding.

When a same sentence is spoken by a native and a non-native speaker, the acoustic features obtained from their speech signals often differ significantly. This change in acoustic observations, in turn, changes the $P_\theta(\mathbf{O} \mid \mathbf{Q})$ (or $y(u,t)$ in case of LF-MMI), during decoding, which accounts for an increase in WER [2, 3, 4]. For the native ASR to perform well on the non-native speech, one possible way is to learn the mapping between native and non-native acoustic features. However, the way the acoustic characteristics in non-native differ from those in native recording could be complex and how it, in turn, impacts the ASR performance may not be straightforward to understand. Instead we hypothesize to learn a corrective mapping between native and non-native pseudo-likelihood vectors (i.e., final layer output of DNN), which directly contribute to the required modification of the acoustic score, and in turn, could improve the WER.

The block diagram of the proposed approach (PLC) is shown in Fig. 1. It involves mainly three components. Each of the blocks are explained in detail below.

Parallel pseudo-likelihood computation: Given non-native and native speech signals for the same sentence, 40-dimensional mel-frequency cepstral coefficients (MFCC) and 100- dimensional iVectors are extracted. The pseudo-likelihood vector, for each MFCC vector is computed using a pre-trained native English model $\mathbf{M}$. In general, the rate of the non-native and native speech signals are different [24]. Hence, to time align the native and non-native pseudo-likelihood vector sequences, we use dynamic time warping (DTW) with Pearson correlation coefficient.

DNN-based pseudo-likelihood correction: Let $\{X_n, Y_n\}_{n=1}^{N}$ be the set of K-dimensional time-aligned non-native (X_n) and native (Y_n) pseudo-likelihood vectors. We propose to map X_n to Y_n by learning a frame-level mapping using a DNN.

Objective Function: MSE could be a typical choice for DNN based PLC mapping. However, it is well known that, in every frame, only a fraction of the elements in the pseudo-likelihood vector contribute to the decoding process [15]. Therefore, in each frame, we consider only the states with top L pseudo-likelihood values in the native pseudo-likelihood vector. We hypothesize that correcting the pseudo-likelihood values corresponding to these L states in the non-native pseudo-likelihood vector is sufficient to obtain reduced WER for non-native speech. To this end, we modify the MSE objective function as follows:

$$\begin{aligned}\mathcal{J}_{topL} = \sum_{n=1}^{N} \big(&\parallel w(X_n, Y_n)^T (Y_n - \hat{Y}_n) \parallel_2^2 + \\ &\parallel (\mathbf{1} - w(X_n, Y_n))^T (\hat{Y}_n - X_n) \parallel_2^2 \big)\end{aligned} \tag{3}$$

where $\hat{Y}_n$ is the estimated pseudo-likelihood vector at the n^{th} frame and $w(X_n, Y_n)$ is the weighting vector function and $\mathbf{1}$ is a K-

dimensional vector of all ones. $w(X_n, Y_n)$ is a vector of ones and zeros, which determines for what states, the MSE between $\hat{Y}_n$ and Y_n is computed. For the remaining states, the MSE between $\hat{Y}_n$ and X_n is reduced, which regularizes $\hat{Y}_n$ to match X_n. Based on how we determine the top L states, we experiment with two kinds of objective function ($Top_L(X,Y)$ and ($Top_L(Y)$)), the details of which are presented in Table 1. L is a hyper parameter and when $L = K$, both the objective functions reduces to MSE between $\hat{Y}_n$ and Y_n only. Thus, MSE is a special case of our proposed objective function.

$Top_L(X,Y)$	$Top_L(Y)$
$w_i(X_n,Y_n) = \begin{cases} 1 & i \in Top_L(X_n,Y_n) \\ 0 & else \end{cases}$	$w_i(X_n,Y_n) = \begin{cases} 1 & i \in Top_L(Y_n) \\ 0 & else \end{cases}$
$Top_L(X_n,Y_n)$: union of the set of indices corresponding to the top L values of X_n and Y_n	$Top_L(Y_n)$: set of indices corresponding to the top L values of Y_n

Table 1. Details of the two objective functions used for DNN mapping, where $w_i(X_n, Y_n)$ is the i^{th} component of the weighting vector $w(X_n, Y_n)$.

4. EXPERIMENT AND RESULTS

The first experiment is conducted to select the optimum parameters corresponding to the minimum WER. These parameters are the value of L and the objective function to be used for training the PLC mapping. In the second experiment, we investigate the robustness of PLC on unseen test sets. Furthermore, the performance of PLC is compared against baseline systems adapted on iTIMIT, MOZ, VOX and iMob for unseen scenarios. We have also investigated the effect of the amount of training data on the performance of PLC against the baseline systems. In the following subsection, details of the source model , baseline along with experiments and results are described.

4.1. Native English model

Base native model (**M**) is trained on 960 hours of Librispeech dataset. The acoustic model is based on sequence-trained time delay neural network (TDNN) with lattice-free maximum mutual information (LF-MMI) objective function. The input features to TDNN are 40-dimensional Mel-frequency cepstral coefficients (MFCC), without cepstral truncation, along with 100-dimensional iVector. The output of TDNN is a 5183-dimensional pseudo-likelihood vector which is used directly for decoding as described in [16]. Further details about the architecture are available in [16].

4.2. Baseline System

As discussed in section 1, for tiny data (5hours), updating the whole network without re-initializing the output layer is found to be beneficial in [9]. Hence, in this paper, we use transfer learning method / model adaptation as described in [9, 11] as the baseline scheme (WA$_S$, S denotes the dataset used for adaptation). The weights of the TDNN network trained on Librispeech is transferred and is further fine-tuned on the non-native datasets with the output layer trained at a relatively higher learning rate. The initial effective learning rate is set to 0.005 and final effective learning rate to 0.0005. Furthermore, the learning rate factor for all the layers is kept 0.25 times the learning rate factor of the output layer. Further implementation details can be found in [9].

4.3. Experiment 1

Parallel Data Preparation: For PLC, we use parallel dataset consisting of a set of utterances spoken by both native English and Indian English speakers. The utterances corresponding to unique 2342 sentences from both TIMIT and iTIMIT datasets are used to create parallel dataset, the steps for which are as follows: 1) From TIMIT dataset utterances corresponding to 2342 unique sentences are sampled from 6300 utterances, such that maximum number of sentences per speaker are obtained and also all the speakers are covered. 2) Further, we split the 2342 utterances into two non-overlapping subsets. Subset-1 contains 1636 (~ 2 hours) utterances from 437 speakers and is used for learning the PLC mapping. Subset-2 contains 706 utterances from 193 speakers and is used for testing the ASR on **M**. 3) We assume that the ASR on non-native English using PLC mapping would benefit, if we cover maximum speaker and accent variability from iTIMIT dataset, in the training set (i.e. subset-1) of parallel data. Hence, corresponding to 1636 sentences, utterances from 63 speakers comprising 12-13 (among 16) randomly selected speakers from each of the 5 groups are chosen. Utterances from remaining speakers corresponding to 706 sentences are used for subset-2.

Experimental Details: The DNN based PLC mapping is learnt based on two objective functions separately. During testing, non-native pseudo-likelihood vector is obtained from the model **M**, which is given as input to PLC to estimate native pseudo-likelihood vector, on which decoding is performed to get the best hypothesis. The DNN network consists of 5183-dimensional input layer, 3 hidden layers and 5183-dimensional output layer. Each hidden layer has 4096 number of units. ReLU is used as activation function along with batch normalization and dropout. In the output layer, linear activation is used. Training of DNN is done using subset-1 of the parallel dataset. On the other hand, the testing is done using iTIMIT part of the subset-2 of the parallel data. For experiments different values of L i.e., $\{1, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 4000, 5000, 5183\}$ have been used. Finally, the objective function and the value of L that yields the minimum WER are chosen for further experiments.

Results & Discussion: The results for experiment 1 are presented in Figure 2. We obtain 12 % and 31% WER on the TIMIT and iTIMIT part of subset-2 respectively, using **M**. The results are indicated by magenta and black dotted line respectively in Figure 2. This indicates that the performance of **M** on the same set of sentences reduces by 19% (absolute) under accent mismatch conditions, i.e., between the Librispeech (training set) and the non-native iTIMIT test set. Furthermore, in Figure 2, WER on iTIMIT part of the subset-2 using PLC is compared across different values of L for both the objective functions. From Figure 2, it can be seen that the performance of both the proposed objective functions at $L = K = 5183$ ($\approx 2^{12}$) is similar to MSE based PLC mapping. It can be observed that WER using MSE based objective function is 14% (absolute) more than that with **M**. This implies that the non-native to native PLC mapping with MSE as objective function is not effective, rather it is detrimental. However, as we start reducing L, WER reduces drastically for both the objective functions. The objective function $Top_L(X,Y)$ yields the minimum WER at $L = 20$ ($\approx 2^4$). Moreover, $Top_L(Y)$ yields the minimum at $L = 1000$ ($\approx 2^{10}$) as indicated by filled circles in the figure. Interestingly, even for $L = 1$, the WER for both the objective functions, is less than WER obtained using **M**. This demonstrates the significance of correcting few values of pseudo-likelihoods in improving the performance of the ASR for non-native test data. The best performance of the proposed approach, i.e., WER 23.9%, is obtained when DNN is optimized by $Top_L(Y)$ objective function for $L = 1000$ ($\approx 2^{10}$).

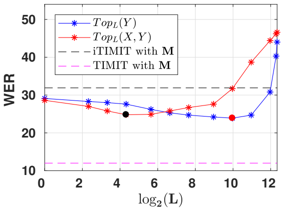

Fig. 2. Comparison of WER for different values of L for two different objective functions. Filled circles indicate the value of L for which the minimum WER is achieved.

4.4. Experiment 2

Data Preparation: In this experiment, we investigate the robustness of the PLC mapping and baseline scheme (adaptation) for unseen recording conditions as well as under varying amount of non-native data. We develop adaptation model on all the datasets, i.e., iTIMIT, MOZ, VOX and iMob as opposed to PLC mapping for which parallel dataset is used. For adaptation on iTIMIT subset-1 of parallel dataset is used. For adaptation on 3 other datasets we randomly select 1636 utterances (~ 2 hours) from each of the corpora such that speaker to gender ratio is maintained. Furthermore, 706 utterances were randomly chosen for testing for each of the 3 other datasets such that the speakers and utterances do not overlap with those in the set used for adaptation.

Experimental Details: We train four different WA models, WA$_{MOZ}$, WA$_{iTIMIT}$, WA$_{VOX}$, WA$_{Mob}$. The performance of these four models is compared with PLC mapping corresponding to the optimal parameters obtained in Experiment 1. We also assess the performance of these models as well as PLC mapping under varying amount of training data . WA and PLC models are, thus, trained using {1636 1000, 750, 500, 250, 100, 75, 50} training utterances. For MOZ, VOX and iMob, the varying amount of adaptation data is obtained by randomly sampling the respective sets of 1636 utterances. For iTIMIT, subset-1 is sampled such that approximately equal number of utterances per group are obtained. Each of these models is tested using the same test sets i.e., 706 utterances from iTIMIT (subset-2) and the test sets obtained from the iMob, VOX and MOZ. To clarify further, for all the schemes, one set is seen and other 3 sets are unseen.

Results & Discussion: Figure 3, presents the results of experiment 2. In Figure 3, each subplot indicates the performance of specific test set using different WA models (i.e., WA$_{iTIMIT}$, WA$_{iMob}$, WA$_{VOX}$, WA$_{MOZ}$) and PLC, when the duration of training/adaptation data is reduced from 120 minutes (1636 utterances) to 4 minutes (50 utterances). Moreover, in each subplot the performance of the test set using **M** is indicated by grey line. It is clear that using **M** the highest WER of 60% is obtained for iMob test set, while the least WER of 31% is obtained for iTIMIT test set. This suggests that the acoustic characteristics of iMob is significantly different from that of Librispeech. Following are the observations from Figure 3: 1) As the amount of data is reduced from 120 minutes to 4 minutes, WER values for all the test sets show increasing trend for all schemes. 2) For the highest amount of data, 1636 utterances (120 minutes), the domain matched case adaptation performs the best with lowest WER, followed by PLC for almost all the cases except for iMob database. This indicates the robustness of proposed PLC scheme. 3) Among all the WA models, WA$_{iTIMIT}$ seems to be the least generalizable model as it shows the worst performance in terms of WER among all other WA models on unseen test case. The possible reason could be highly mismatched recording conditions of iTIMIT (i.e., clean lab environment) as opposed to other datasets. 4) On the contrary PLC which is also trained on clean iTIMIT set, shows comparable performance to that of WA$_{MOZ}$ on the unseen test cases. Furthermore, for lower number of training data points, PLC outperforms WA$_{MOZ}$. This indicates that PLC is robust to highly mismatched recording conditions. 5) For very small amount of data, i.e., as low as 4 minutes, PLC has the least WER compared to other schemes for all test sets. For iMob test set, PLC always performs better than **M**, even for as low as 4 minutes of train data. In addition, 6 - 8 minutes of training data is sufficient for PLC, to surpass the performance of the native model **M** for not only iTIMIT set (matched case) but also for all other test sets. 6) The performance of PLC saturates beyond 60 minutes of training data across all tests, which indicates that the proposed PLC approach doesn't over-fit the training data. However, WA models shows a decreasing WER trend with increase with training data for matched test case, which can be indicative of over fitting.

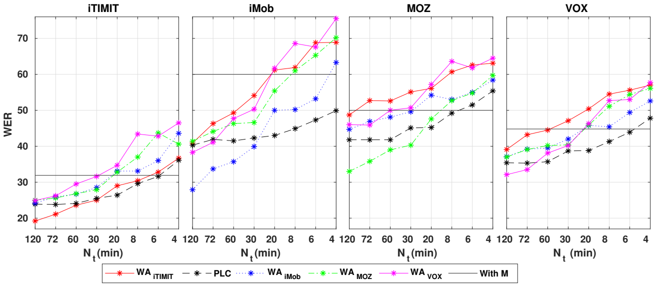

Fig. 3. Comparison of WER for amount of training utterances (N_t) for different databases. The title of the plot shows the test-set. WA$_m$ indicates the adapted model using database m.

5. CONCLUSIONS

In this paper, we experimented with DNN based PLC mapping to improve the performance of ASR for Indian English speakers with varied mother tongue. We proposed novel objective function to learn the parameters, that optimizes only top L values of the pseudo-likelihood vector. The experiments reveal that optimizing PLC mapping using standard MSE objective function is detrimental for the non-native ASR performance. On the contrary, the proposed objective function showed significant improvement in WER compared to native model performance, as L value is reduced from K, indicating significance of using top L for PLC mapping. With limited amount (2 hours) of training data, the best performance gives 7% relative improvement over native ASR system (**M**). Even though the relative improvement is 5% less than the baseline, PLC generalizes well across different unseen datasets. For very small amount of data, i.e., as low as 4 minutes, PLC has the least WER compared to other schemes for all test sets. From the results, it can be clearly seen that PLC shows consistent improvements over seen and unseen test sets compared to WA models obtained on different datasets for even small amount of training data which is 20 minutes and less. Thus using top L PLC can generalize well in low resource conditions. In future, we will like to experiment on the robustness of this approach for unseen accents. Furthermore, we would like to investigate if the performance of PLC is specific to the choice of data used in the experiments or not and moreover, how this approach can be applied to a new accent in low resource scenarios.

6. REFERENCES

[1] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778.

[2] Dirk Van Compernolle, "Speech recognition by goats, wolves, sheep and non-natives," *Multi-Lingual Interoperability in Speech Technology*, p. 1, 2000.

[3] Antoine Raux and Maxine Eskenazi, "Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges," in *InSTIL/ICALL Symposium*, 2004.

[4] Kacper Radzikowski, Robert Nowak, Le Wang, and Osamu Yoshie, "Dual supervised learning for non-native speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 3, 2019.

[5] Zhirong Wang, Tanja Schultz, and Alex Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP).*, 2003, vol. 1, pp. 540–543.

[6] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Daniel Jurafsky, Rebecca Starr, and Su-Youn Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 217–220.

[7] Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain, "Automatic speech recognition of multiple accented English data," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Tien-Ping Tan, "Merging of native and non-native speech for low-resource accented ASR," in *International Conference on Statistical Language and Speech Processing*, 2015, pp. 255–266.

[9] Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Jhu Kaldi system for Arabic MGB-3 ASR challenge using Diarization, audio-transcript alignment and transfer learning," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 346–352.

[10] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[11] Pegah Ghahremani, Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur, "Investigation of Transfer learning for ASR using LF-MMI trained neural networks," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 279–286.

[12] Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, and Yushi Aono, "Domain adaptation of DNN acoustic models using knowledge distillation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5185–5189.

[13] Jinyu Li, Michael L Seltzer, Xi Wang, Rui Zhao, and Yifan Gong, "Large-scale domain adaptation via teacher-student learning," *arXiv preprint arXiv:1708.05466*, 2017.

[14] Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 250–257.

[15] Dong Yu and Li Deng, *Automatic Speech Recognition – A Deep Learning Approach*, Springer, 2016.

[16] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[18] John S Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[19] "Voxforge," `http://voxforge.org/`, [Online; accessed 25-June-2019].

[20] "Common voice speech corpous," `https://voice.mozilla.org/`, [Online; accessed 25-June-2019].

[21] Chiranjeevi Yarra, Ritu Aggarwal, Avni Rajpal, and P.K. Ghosh, "Indic TIMIT and Indic English lexicon: A speech database of indian speakers using TIMIT stimuli and a lexicon from their mispronunciations," accepted in Oriental COCOSDA 2019. [Currently available at: https://spire.ee.iisc.ac.in/spire/conferences.php].

[22] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *Interspeech*, 2013, vol. 2013, pp. 2345–2349.

[23] Naoyuki Kanda, Yusuke Fujita, and Kenji Nagamatsu, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level kullback-leibler divergence," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 69–76.

[24] Melissa M Baese-Berk and Tuuli H Morrill, "Speaking rate consistency in native and non-native speakers of English," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. EL223–EL228, 2015.