PSEUDO LIKELIHOOD CORRECTION TECHNIQUE FOR LOW RESOURCE ACCENTED ASR

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

SPIRE LAB
Electrical Engineering
Indian Institute of Science (IISc), Bengaluru, India

Poster Session: TH3.PC: Speech Recognition: Adaptation
Thursday, 7 May, 2020, 16:30 - 18:30

Plan

- Introduction
- Proposed Pseudo Likelihood Correction (PLC) Approach
- Objective Function for PLC
- Experiments & Results
- Conclusion & Future Work
- References

Introduction

- ASR trained on native English performs poorly on non-native English
- Primary Factor: **high confusion in posteriors** obtained from native acoustic model due to unseen accent variations
- Performance of current methods are **limited by the availability of data**
- **Proposed Approach:**
  - With ~2 hours parallel data learn DNN-based pseudo-likelihood correction (PLC) mapping
  - Non-native pseudo-likelihood vector is mapped to match its native counterpart

Background

- The fundamental equation of ASR:

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

$\mathbf{O}$: sequence of acoustic feature vectors

$\mathbf{W}$: sequence of words

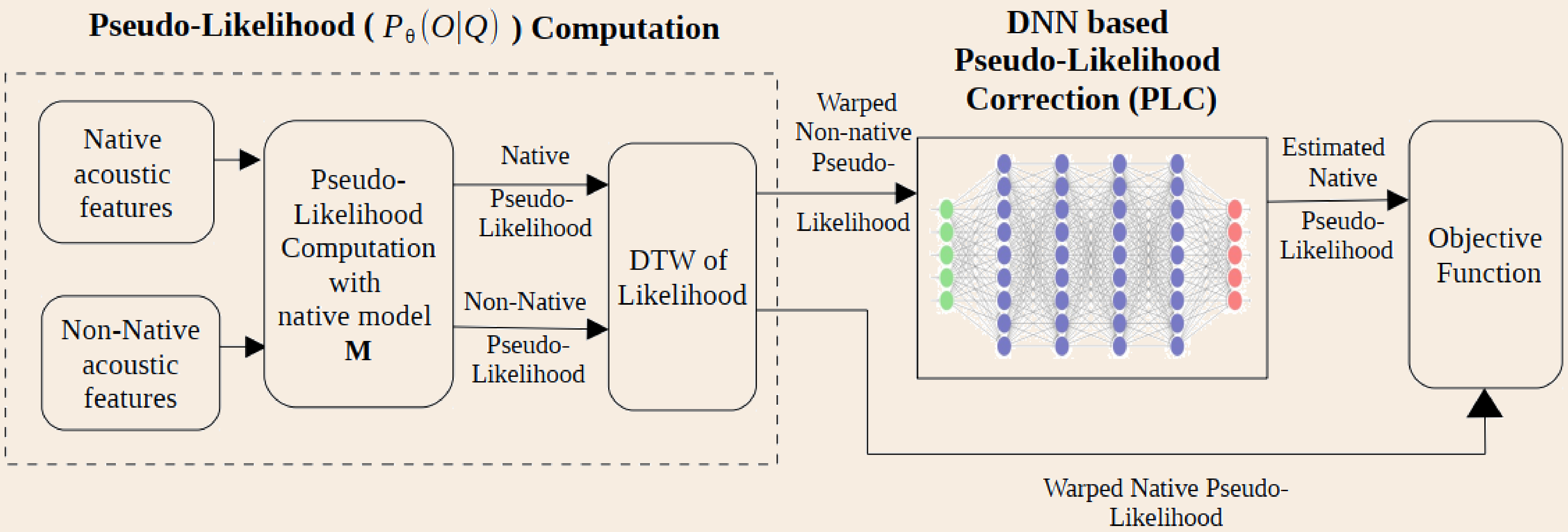
$\mathbf{Q}$: sequence of states

θ: set of parameters of the HMM-DNN model, estimated using LF-MMI

- $P_\theta(\mathbf{O} \mid \mathbf{Q})$ (interpreted as pseudo-likelihood vector) is the DNN output without softmax activation and is directly used as acoustic score for decoding [1]

[1] Daniel Povey et al., “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in Interspeech, 2016, pp.2751–2755.

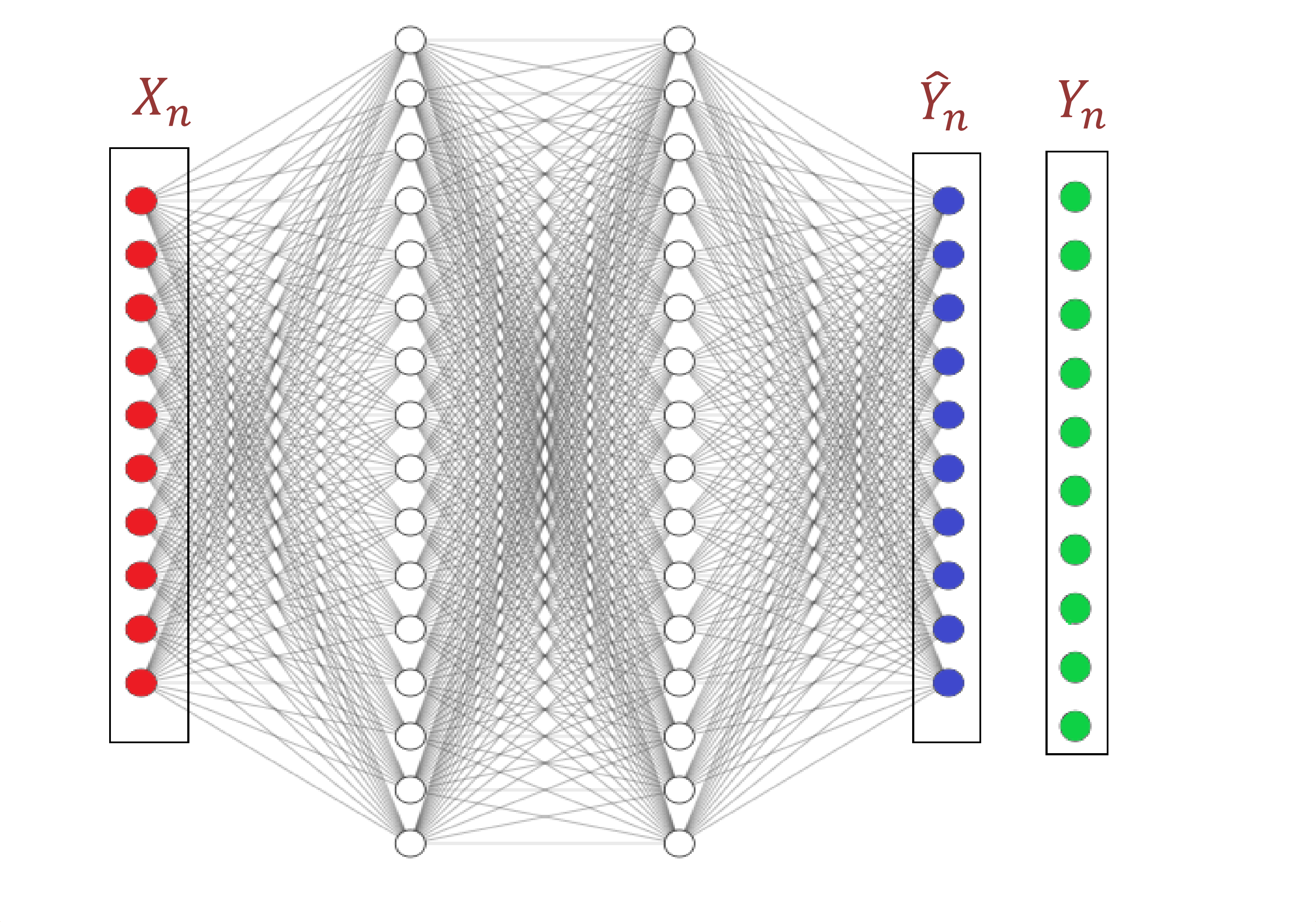
Block Diagram



Acoustic features: 40-d MFCC+ 100d ivector

Native English ASR Model(M): Nnet3-chain TDNN model trained on Librispeech

Objective Function



Mean Squared Error(MSE):

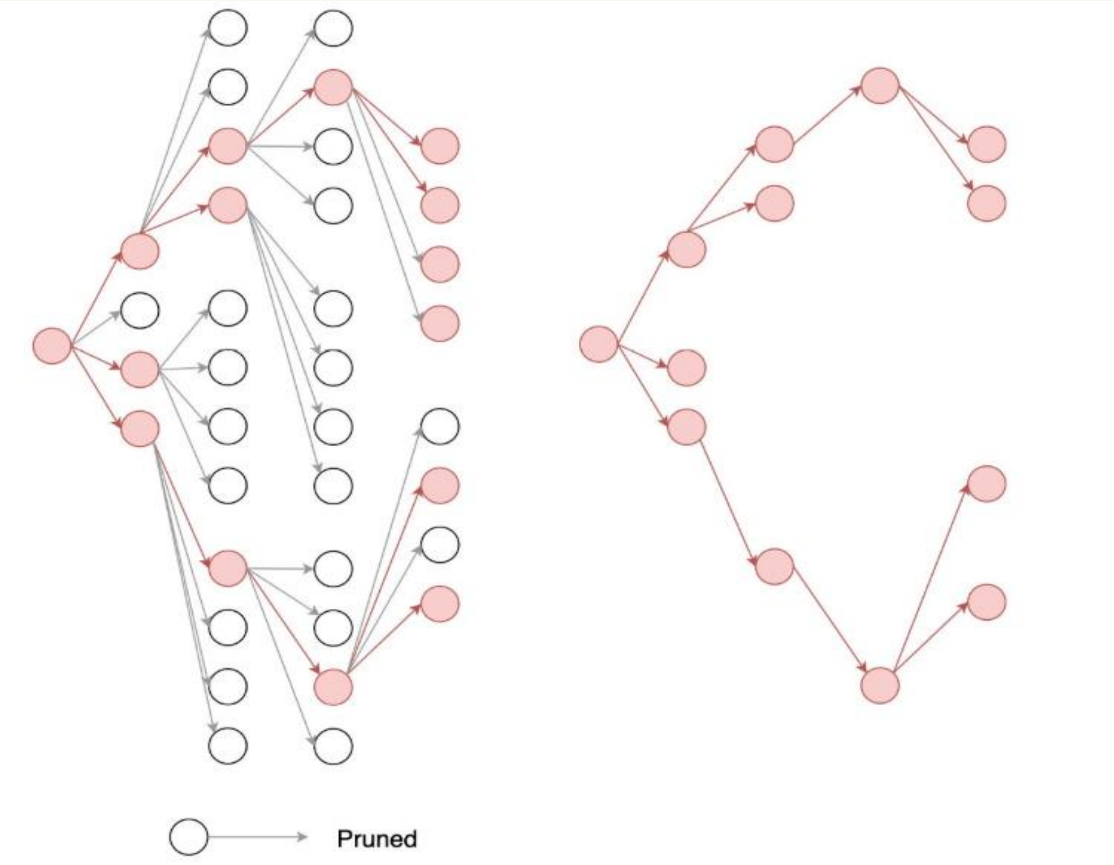
$$\mathcal{J} = \sum_{n=1}^{N} (\| (Y_n - \hat{Y}_n) \|_2^2)$$

X_n: K-d Warped Non-Native pseudo-likelihood vector

$\widehat{Y_n}$: K-d Estimated pseudo-likelihood vector

Y_n: K-d Warped Native pseudo-likelihood vector

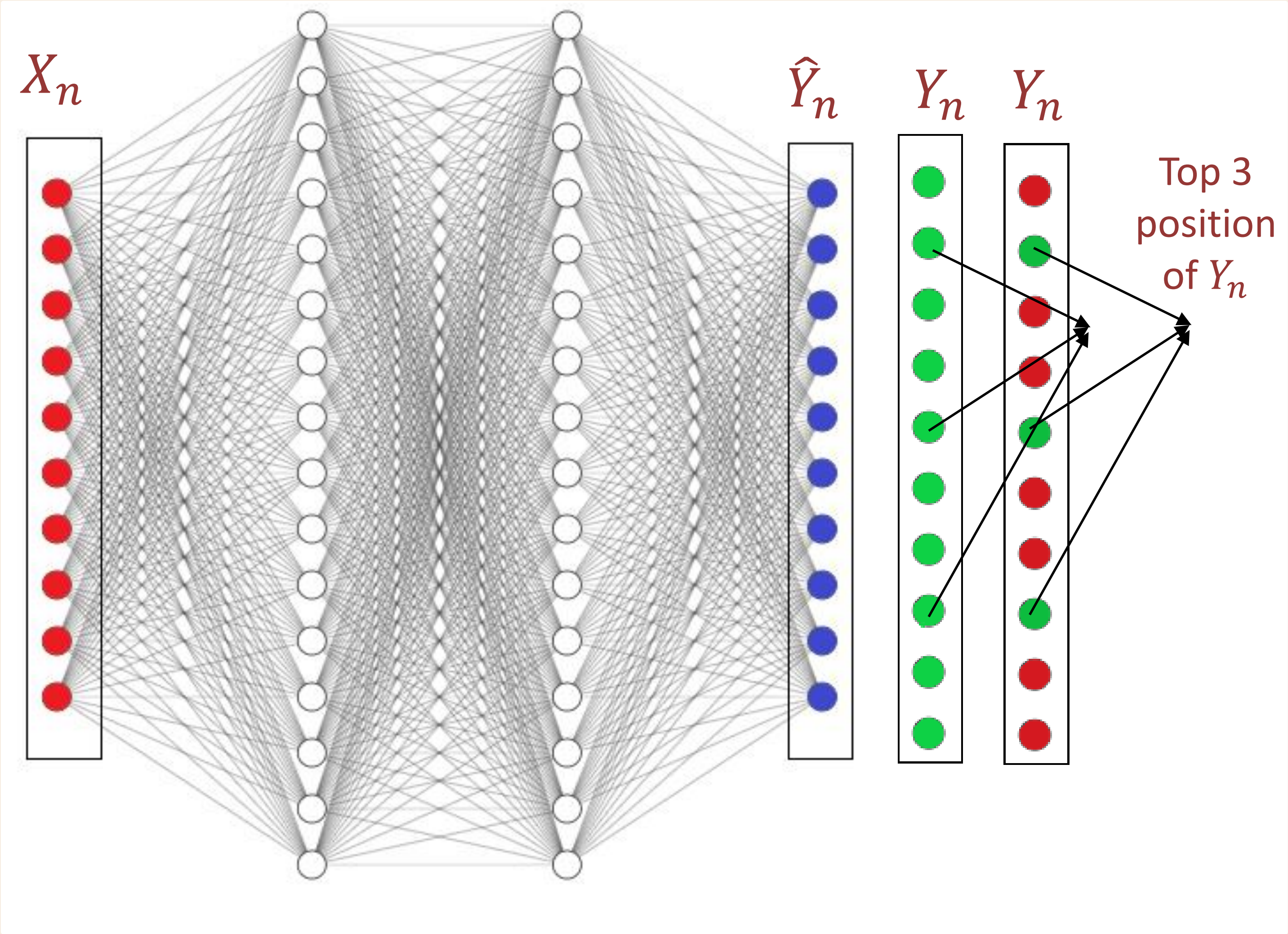
N: Total number of training examples



- In ASR decoding process only top few state score values per frame contribute in obtaining optimal hypothesis [2]
- **Proposed objective function**: Considers only top $\mathbf{L}$ values of pseudo-likelihood vector

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

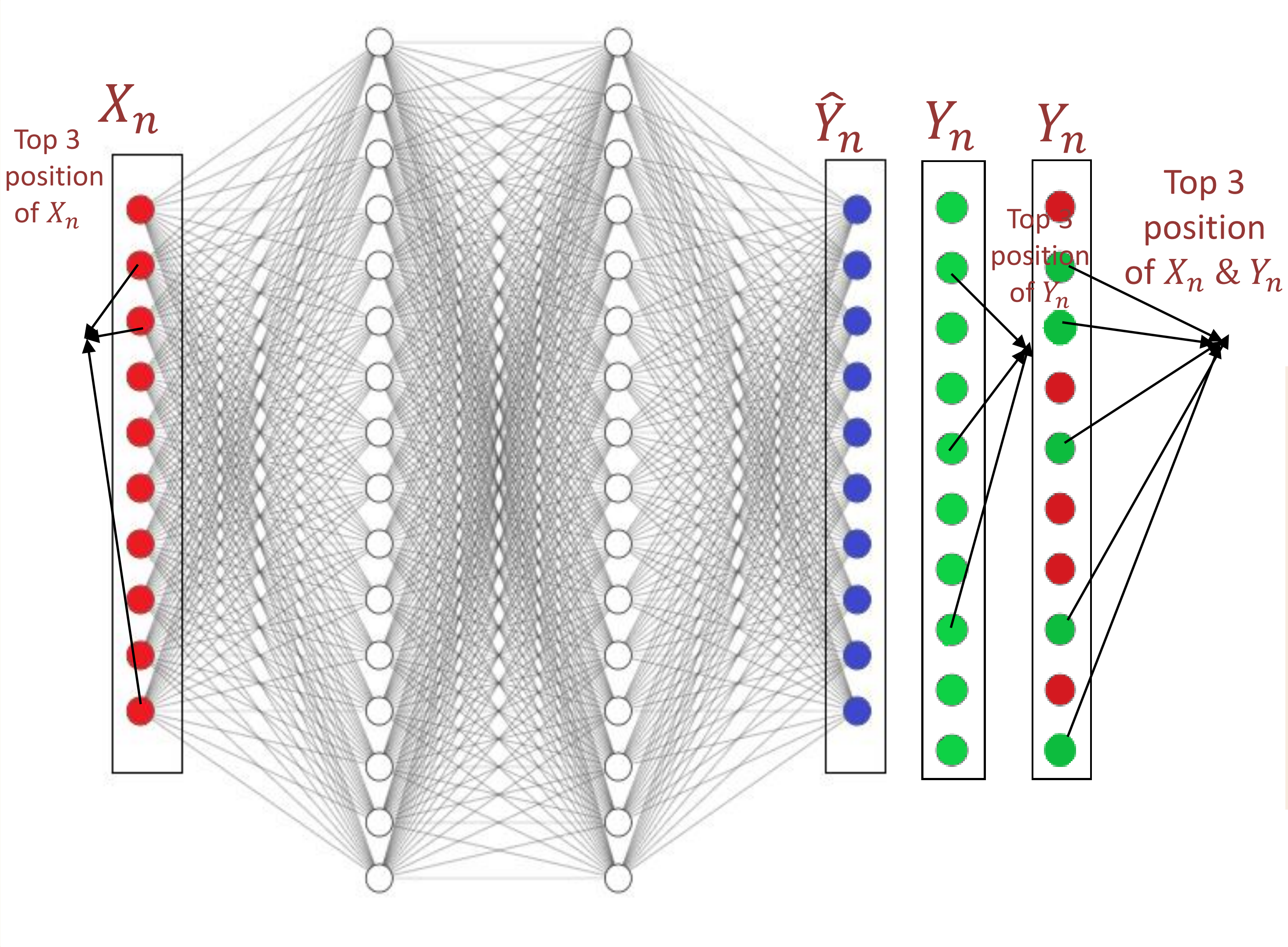
PLC: Objective Function 1



$$\mathcal{J}_{topL} = \sum_{n=1}^{N} \left(\| w(X_n, Y_n)^T (Y_n - \hat{Y}_n) \|_2^2 + \| (\mathbf{1} - w(X_n, Y_n))^T (\hat{Y}_n - X_n) \|_2^2 \right)$$

Objective Function: $Top_L(Y)$
$w_i(X_n, Y_n) = \begin{cases} 1 & i \in Top_L(Y_n) \\ 0 & else \end{cases}$
$Top_L(Y_n)$: set of indices corresponding to the top L values of Y_n $w_i(X_n, Y_n)$: i^{th} component of the weighting vector $w(X_n, Y_n)$

PLC: Objective Function 2



$$\mathcal{J}_{topL} = \sum_{n=1}^{N} \Big(\| w(X_n, Y_n)^T (Y_n - \hat{Y}_n) \|_2^2 + \| (\mathbf{1} - w(X_n, Y_n))^T (\hat{Y}_n - X_n) \|_2^2 \Big)$$

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

At L=K (dimension of pseudo-likelihood vectors), both objective functions, reduces to MSE

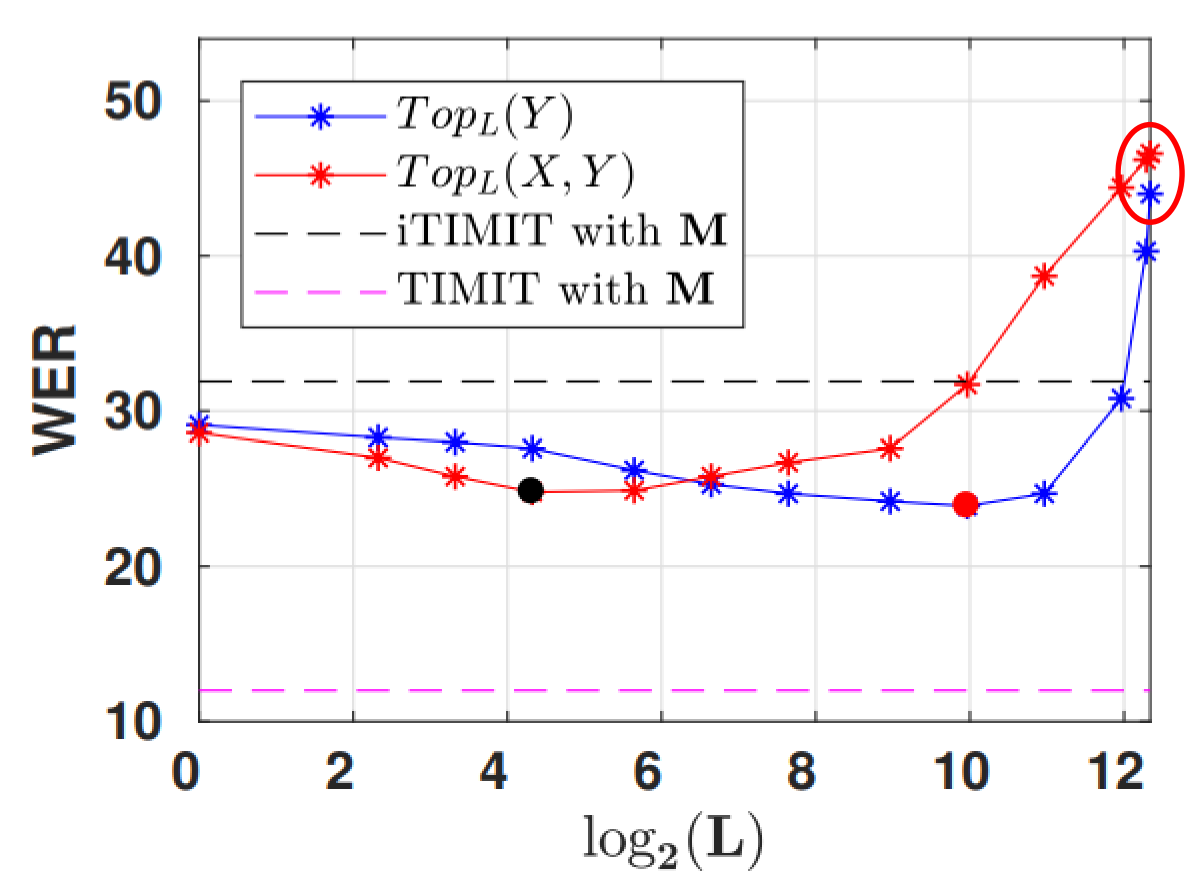
Experimental Setup

Technique	Architecture Details	Training/ Adaptation Set	#of Training utterance	Test Set	#of Test utterance
PLC	3-layer DNN, with 4096 hidden units	Parallel set from iTIMIT and TIMIT dataset	1636 (~2 hours)	iTIMIT, iMob, MOZ, VOX	706
Baseline (WA_m)	weights of **M** fine-tuned on the non-native datasets. [3]	Adaptation using m dataset	1636 (~2 hours)	iTIMIT, iMob, MOZ, VOX	706
*m indicates datasets used for adaptation which are iTIMIT, iMob, Common Voice (MOZ), Voxforge (VOX)					

- **iTIMIT**: Indian English dataset of 80 speakers collected in our **laboratory environment**. Each speaker records 2342 sentences from the TIMIT corpus [4].
- **iMob**: 100 hours of Indian English dataset of 827 speakers collected by us through mobile application with the help of industry partner.
- **Common Voice** (MOZ) and **Voxforge** (VOX) are **publicly available datasets** from which the Indian English voice samples are considered for our experiment.

[3] Pegah Ghahremani, et al. , "Investigation of Transfer learning for ASR using LF-MMI trained neural networks," in Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 279–286
[4] Chiranjeevi Yarra, et al. "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.

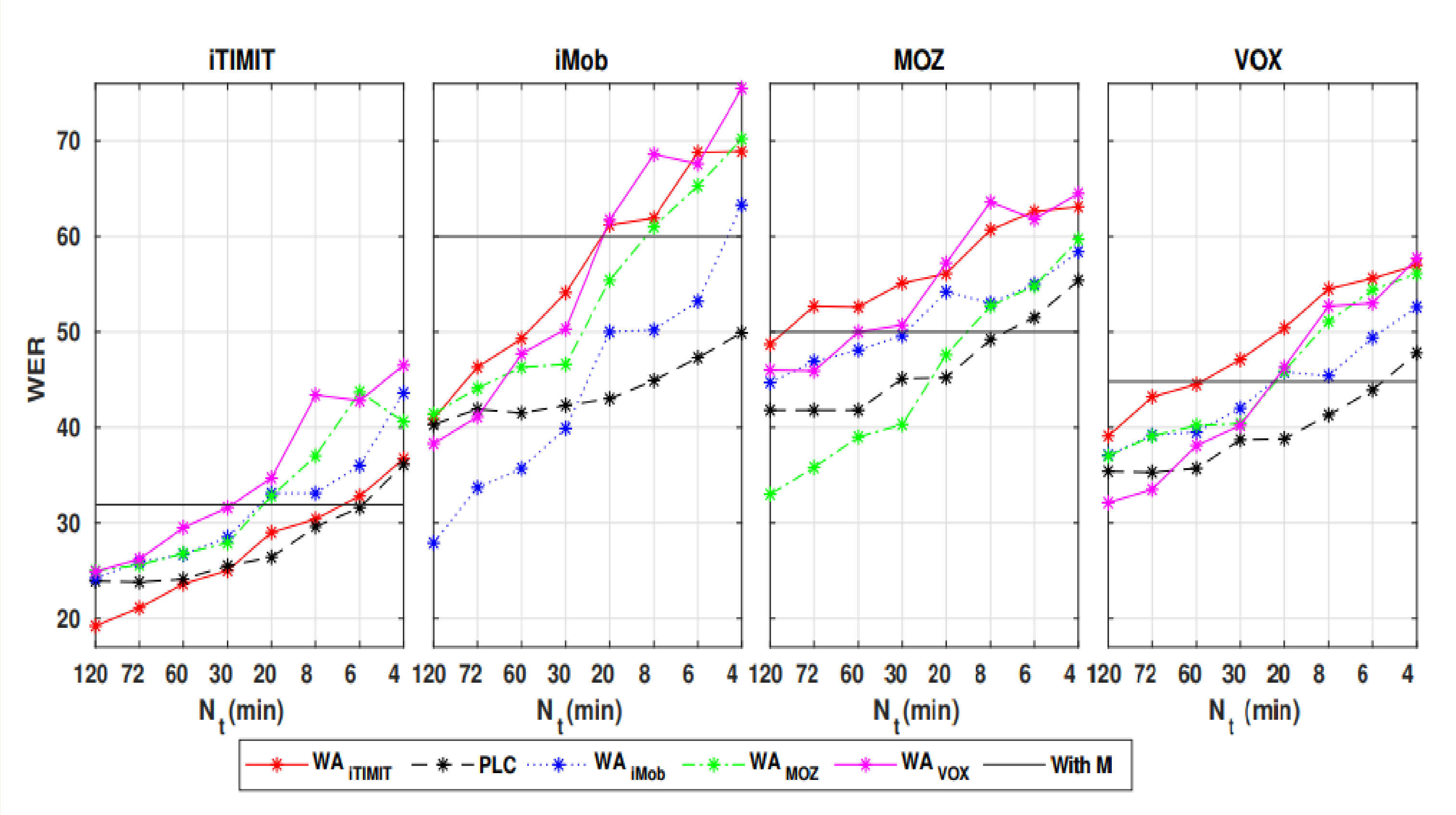
Experiment 1



Native English ASR Model(M): Nnet3-chain TDNN model trained on Librispeech

- **L** is varied from *1-5183*
- 5183: dimension of pseudo-likelihood vectors
- WER (TIMIT): 12 %
- WER (iTIMIT): 31%
- As **L** reduces WER reduces
- **L**=5183, both objective functions, reduces to MSE
- At **L**=5183 (all states considered), WER > WER(iTIMIT) with **M**
- Best Performance is obtained for *$Top_L(Y)$* at **L**=1000 , with WER: 23.9%
- Thus PLC shows **~ 7% improvement** compared to **M** for iTIMIT dataset.

Experiment 2



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.

- N_t varied from 120 min-4 min
- WER least for PLC for unseen cases
- WER for PLC is the lowest for all cases, for N_t <=20 min
- PLC is robust to highly mismatched recording conditions
- WA_{iTIMIT} is the least generalizable with maximum WER for all unseen cases
- WER of PLC saturates for N_t > 60 min for all test sets

Conclusion & Future Work

- DNN based PLC mapping **optimized over top L values** of the pseudo-likelihood vector is proposed.
- The best performing system trained with ~ 2 hours of data **yields 7% improvement** over native ASR system.
- PLC is found **robust to highly mismatched recording conditions.**
- PLC has the least WER compared to other schemes for all test sets with training data <=20 min, indicating **generalizing capability in low resource conditions.**
- In future, we will like to investigate the robustness of PLC for unseen accents and on the choice of dataset used.

References

- 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.
- Dong Yu and Li Deng, Automatic Speech Recognition – A Deep Learning Approach, Springer, 2016.
- 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.
- 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

Acknowledgement

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

THANK YOU

Questions/Suggestions?

Write to us at spirelab.ee@iisc.ac.in