Acoustic models for speech recognition in Reading Miscue detection

MTP Presentation

Shreeharsha B S
EE1 18307R002

Motivation

Literacy is an important measure of prosperity and also critical to the well being of an individual and his/her community.

ASER (Annual Status of Education Report) survey[1]:

27.2% of class VIII students sampled unable to read a text that is meant for a class II student[2].

The ASER annual survey is an important tool to guide education interventions that is widely respected by governments and NGOs.

Goal: Build a reliable and scalable system for school level reading skill assessment based on automatic speech recognition (ASR) and fluency detection.

Introduction

Challenges:

1) Scarcity of labeled target speech
2) High degree of variability:
   a) Variation in speaking and reading abilities due to possible L2 context
   b) Variation in accents due to regional differences and other influences
3) Noisy environments

Helpful factors: Reading of known text

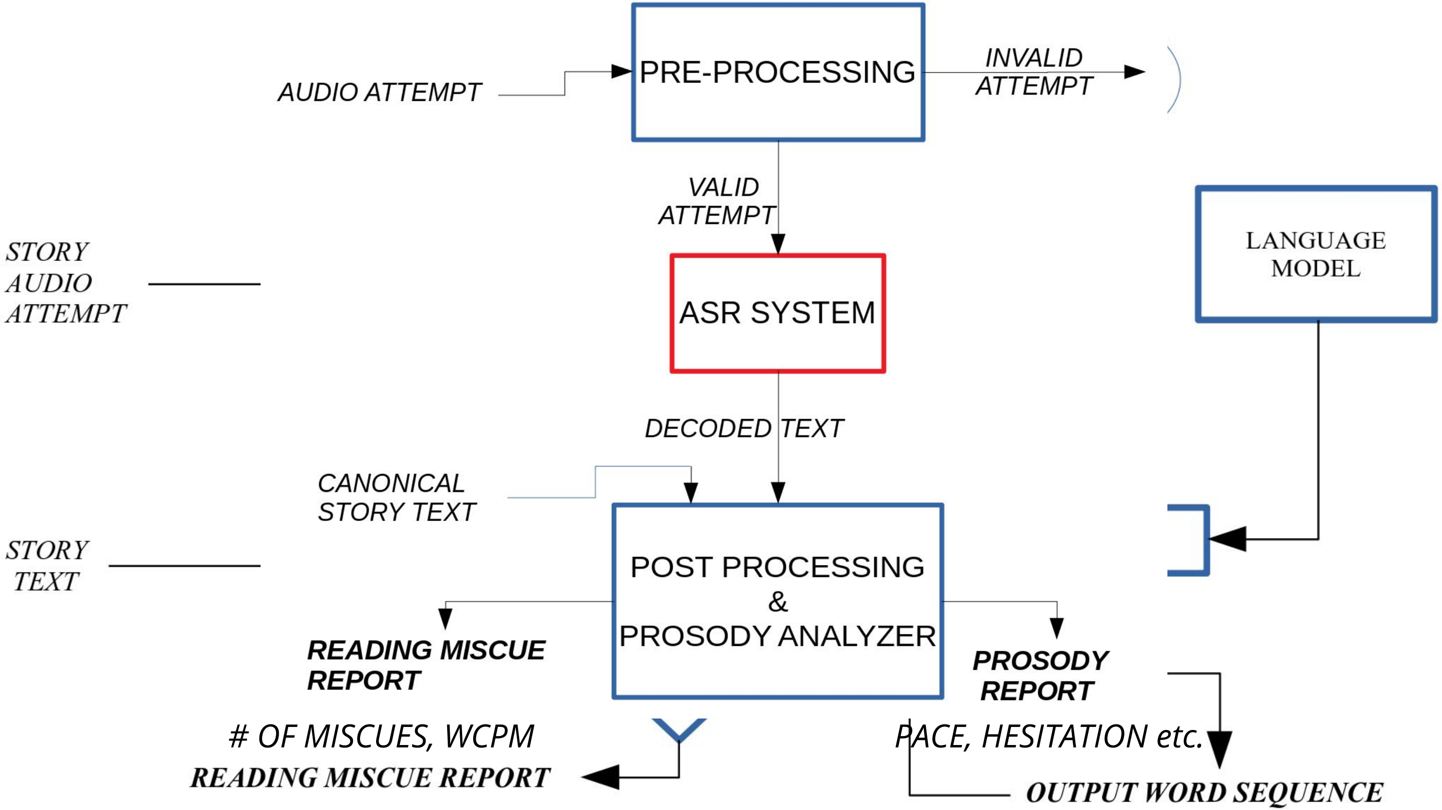
Approach:

Use a baseline system trained on more widely available adult speech in the target language

Investigate data augmentation and transfer learning using available target domain data

Overall System for Automatic Reading Assessment

ASR System



ASR System - Kaldi baseline Acoustic models

13 TDNN layers + 2 linear layers + 2 output layers model trained on MMI and cross entropy loss functions

Use two different output layers (cross entropy and MMI loss) while training. Only MMI layer used in decoding

```
relu-batchnorm-dropout-layer name=tdnn1 dim=1024
tdnnf-layer name=tdnnf2  dim=1024 time-stride=1
tdnnf-layer name=tdnnf3  dim=1024 time-stride=1
tdnnf-layer name=tdnnf4  dim=1024 time-stride=1
tdnnf-layer name=tdnnf5  dim=1024 time-stride=0
tdnnf-layer name=tdnnf6  dim=1024 time-stride=3
tdnnf-layer name=tdnnf7  dim=1024 time-stride=3
tdnnf-layer name=tdnnf8  dim=1024 time-stride=3
tdnnf-layer name=tdnnf9  dim=1024 time-stride=3
tdnnf-layer name=tdnnf10  dim=1024 time-stride=3
tdnnf-layer name=tdnnf11  dim=1024 time-stride=3
tdnnf-layer name=tdnnf12  dim=1024 time-stride=3
tdnnf-layer name=tdnnf13  dim=1024 time-stride=3
linear-component name=prefinal-l dim=192
prefinal-layer name=prefinal-chain input=prefinal-l  big-dim=1024 small-dim=192
output-layer name=output dim=3080
prefinal-layer name=prefinal-xent input=prefinal-l  big-dim=1024 small-dim=192
output-layer name=output-xent dim=3080
```

Baseline system and Tasks

Train a Hindi baseline model and an Indian English baseline model using Adult speech in Hindi and Indian English respectively

Summary of the IITM datasets

Dataset	# of Utterances	# of Unique speakers	Duration (min)
IITM Hindi train	27131	418	2400
IITM Indian English train	55330	598	4800

Baseline model recipe and Data: IITM Hindi and English ASR challenge[3]

Hindi Task: Transfer learning & data augmentation experiments on Hindi baseline model using ASER Hindi data[4]

English Task: Similar experiments on Indian English baseline model using children's English data

Manual phone mappings made between retraining data lexicons and IITM baseline lexicon

Hindi task

Recordings of ASER survey conducted by trained volunteers to manually evaluate literacy levels of students

Task: Automate the process of reading evaluation through ASR on audio recordings of tests

Datasets:

Summary of the ASER datasets

Dataset	# of recordings	# of Unique speakers	Duration (min)
2012 UP	1488	915	697
2012 RJ	1478	1007	634
2016 CG	418	251	239
2016 JH	375	232	218
2016 MH	145	84	70
2016 RJ	281	170	144
2016 UK	62	40	33

Hindi task: ASER survey sample test


सैम्पल

असर के बुनियादी पढ़ने की जाँच सामग्री: हिन्दी

कक्षा II स्तर का पाठ

सावन का महीना था। आसमान में बहुत काले-काले बादल छाए थे। ठंडी-ठंडी हवा चल रही थी। मुझे झूला झूलने का मन किया। बड़े भैया एक मोटी सी रस्सी लेकर बाहर आए। भैया ने रस्सी को पेड़ से लटकाकर झूला बनाया। सब ने मिलकर खूब झूला झूला। बाकी बच्चे भी आकर मज़े से झूलने लगे। झूलते-झूलते रात हो गई।

नोटः यह पाठ भारत में सारी कक्षा I और II की पाठ्य पुस्तकों का विश्लेषण करके तैयार किया गया है।

यह बुनियादी पढ़ने की जाँच का एक सैम्पल है।

पढ़ने की जाँच की सामग्री सभी भारतीय भाषाओं में उपलब्ध है।
www.asercentre.org देखें, ई-मेल: contact@asercentre.org

कक्षा I स्तर का पाठ

बग़ीचे में एक पेड़ है।
पेड़ पर एक तोता रहता है।
तोते का रंग हरा है।
वह लाल टमाटर खाता है।

अक्षर

ल प स
क ग
ड ब म
ट झ

सामान्य आसान शब्द

लाल दूध
पैर
तेल किला
मोर
जूता मौका

अक्षर/शब्द के लिए: बच्चे से कोई 5 पढ़ने को कहें, कम से कम 4 सही होने चाहिए।

An ASER sample test[2] containing letter, words, paragraphs and stories.

In this work, only the stories and paragraphs are used.

Manual Transcription

Audacity sentence level Label Track: 1560_08_aser_up_HI-S1-P_2

0.000000	5.096792	SIL ON माँ ने हलवा बनाया
5.096792	7.340000	वह बहुत मीठा था ON
7.340000	11.900000	उसे उसे सोनी ने SIL खाया SIL
11.900000	16.496487	खाने SIL के बाद SIL SIL वहा सो गई
16.496487	21.710000	ON IR ON SIL ON

Canonical text: उसे सोनी ने खाया खाने के बाद वह सो गई

Training transcription: उसे उसे सोनी ने SIL खाया SIL खाने SIL के बाद SIL वहा सो गई ON IR ON SIL ON

True transcription: उसे उसे सोनी ने खाने के बाद वहा सो गई <u>तुम्हारा पहला ही सेट नही हो पाया, जल्दी जल्दी करो, हो गया</u>?

Miscues: ICCCCCCSCC ;**Miscue rate** = (I+S+D)/(# of words in story) b/w **Training transcript** and **Canonical**

IR: Irrelevant speech **BR**: Breathing in/out **MB**: Undecipherable mumbling **ON**: Other noise **FP**: Filled pauses
SIL: Silences > 200 ms

ASER 2012 Data Splits

Train, valid and test splits made from 2012 data. 12 unique Hindi stories

Summary of the ASER datasets

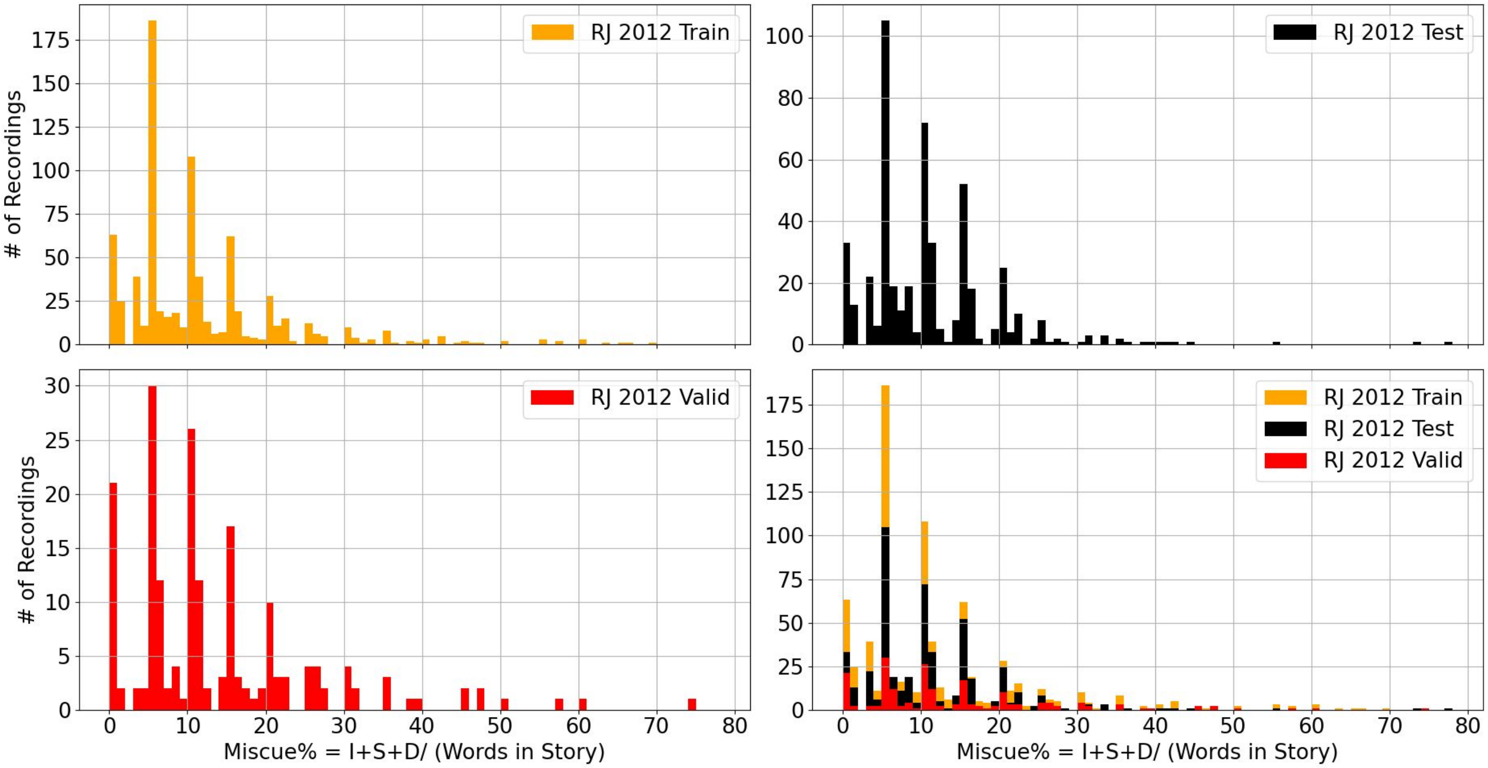
Dataset	# of recordings	# of Unique speakers	Duration (min)
2012 UP	1488	915	697
2012 RJ	1478	1007	634
2016 CG	418	251	239
2016 JH	375	232	218
2016 MH	145	84	70
2016 RJ	281	170	144
2016 UK	62	40	33

Dataset	min)
2012 UP train	
2012 UP test	
2012 UP valid	
2012 RJ train	
2012 RJ test	
2012 RJ valid	

2012 UP+RJ Train set combined (12 hrs) split at sentence level, used for retraining baseline along with Hindi data from campus school

No speaker overlap between train, valid and test splits

2012 splits Miscue rate distribution



Many recordings have miscue rates b/w 0-10%

ASER 2016 Data splits

2016 set used only for decoding and testing.

Two subsets: With and without story overlap with 2012 data.

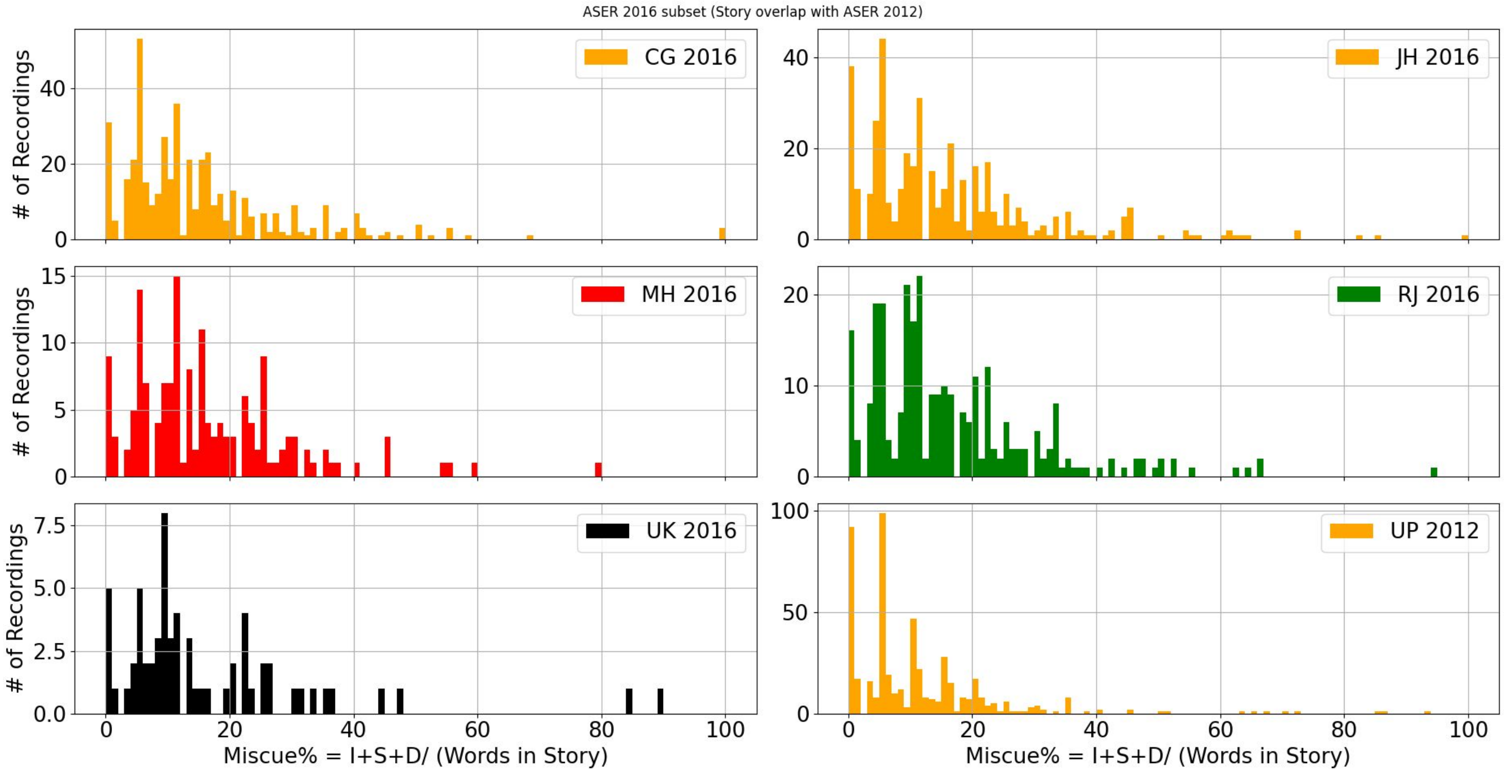
No story overlap (8 new unique Hindi stories) further split into valid and test.

Summary of the 2016 ASER Hindi datasets

Dataset	# of Recordings	# of Unique speakers	Duration (min)
2016 no story overlap valid	482	317	252
2016 no story overlap test	333	220	192
2016 with story overlap	459	289	256

More challenging dataset because of noisier conditions, children are from 5 different states (CG, JH, MH, RJ, UK) and make more mistakes while reading.

2016 splits Miscue rate distribution



Many recordings have miscue rates b/w 10-20%

Data Augmentation

Data augmentation → Apply certain transforms on the training data to:

1) Enhance amount of training data
2) Groom the model towards certain test scenarios

Augmentation techniques:

VTLP (Vocal tract length perturbation) [5]

SpecAugment[6] and SpecSwap[7]

Speed perturbation, Tempo perturbation, VTLP examined in kaldi[8]

Pitch perturbation [9] and Noise augmentation

A mix of augmentation techniques are used in baseline model and during transfer learning depending on the scenario of interest

Data Augmentation procedure

Two techniques used during **baseline model training**:

1) Vocal tract length perturbation (VTLP) warping
2) Speed perturbation at 1.1x and 1.2x

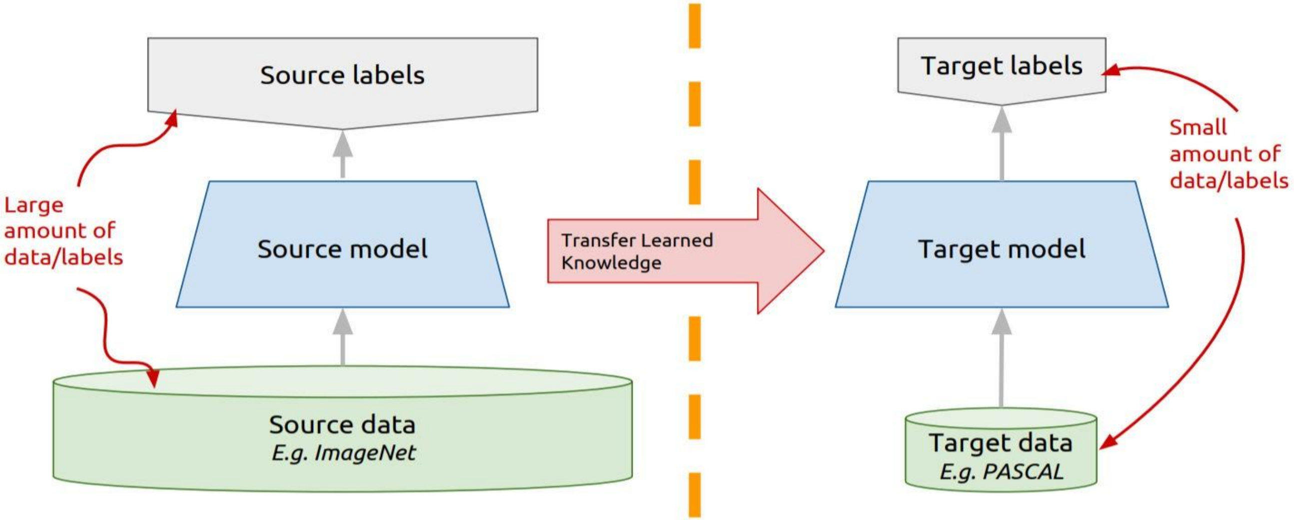
During **retraining on ASER 2012 data**:

1) Overlaying noise sources (IITB exam recordings, wind, babble, rain, traffic etc) at various SNR. (2x versions of the retraining acoustic data (original+noisy))
2) Original and noisy used for SpecAugmentation once (4 versions)
3) Speed perturbation of above 4 versions at 0.9x, 1.0x, 1.1x (12 versions)
4) Pitch perturbation of the 4 versions from 2) by upshifting and downshifting at ~0.9x and ~1.1x pitch of each individual recording (8 versions)

In total, there would be 20 versions (12 hrs → 240 hrs) of the retraining acoustic data available

Transfer learning

Use source model trained on large amount of data and adapt it to target data that is more scarce



A general block diagram of transfer learning[9]

Transfer learning parameters tuned

Regularization:

1) L2-regularization: adding $-0.5c*|y|_2$ to the loss function, where y is the output of any node in the MMI 'output' layer
2) xent-regularization: scales xent (cross entropy) layer output (output/xent-regularize)

# of epochs:

# of passes of retraining data through the model. Determines total # of training iterations along with mini-batch size (64).

Global initial and final effective learning rates:

Learning rate at each iteration. Starts at the initial learning rate on first iteration and decreases at each iteration until it reaches the final learning rate

In all experiments, the final effective learning rate is 1/10th the initial learning rate.

Transfer learning parameters tuned

Differential learning rates:

Train different layers of the baseline model at different rates.

Represented like so: ”5(4)-0(9)-0.625(2)*1e-6”: first 4 TDNN layers have initial learning rates 5e-6, the next 9 layers are frozen and the final 2 output layers have initial learning rate 0.625e-6.

The middle layers of the network are frozen and the top and bottom layers retrained[10]:

1) Acoustic variability in the features is captured by the layers near input of the model
2) Pronunciation variability is seen only at the layers near the output of the model.

Chunk-width (C/chkwidth):

Each training mini-batch consists of an 64 x C x 140 tensor. Since the loss function can be computed on any sequence of frames, the network can be made to learn text contexts of length C (along with the acoustic characteristics of the speakers) during retraining.

Evaluation

Word Error Rate (WER)

WER calculated using a tri-gram LM trained on canonical stories' text with a unigram garbage model (words from retrain + valid transcript that occurred at least twice).

Other labels (IR, FP, SIL etc) removed from GT (Ground truth) text before computing WER

F-score of Detected Correct words

Precision (P): Of the total number of correct words identified by the ASR system how many were actually correctly spoken by the child in the GT.

Recall (R): Of the total number of correctly spoken words by the child according to the GT, how many were identified by the ASR system.

Here, a correct word is a word present in the canonical text and correctly spoken by the child. The F-score is then $(2 * P * R)/(P + R)$

Hindi task results

Experiments involving freezing of various layers/blocks of the TDNN

Retraining parameter setting on Hindi Baseline model	Diagnostic 2012 Train UP+RJ MMI loss	2012 UP+RJ Validation MMI loss
5(13)-2.5(2) * 1e-6	0.443576	0.368796
5(1)-0(12)-0.625(2) * 1e-6	0.4443	0.369501
5(2)-0(9)-0.625(4) * 1e-6	0.444404	0.369544
5(3)-0(8)-0.625(4) * 1e-6	0.444407	0.369603
5(4)-0(9)-0.625(2) * 1e-6	0.444351	0.369564
5(4)-0(6)-0.625(5) * 1e-6	0.44442	0.36959

Best results with freezing middle 8 layers; used for all further experiments

ASER 2012 Decoding results

High Proficiency Recordings (**HPR**): miscue rate <=20%

LPR Proficiency Recordings (**LPR**): miscue rate >20%

Improvements obtained on 2012 Validation (UP+RJ) data

LM: 3 gram 2012 canonical ASER stories

GM: UP train+valid words (count>=2)

Acoustic Model	HPR WER%	LPR WER%	HPR F-score (P, R)	LPR F-score (P, R)	HPR lmwt, wip	LPR lmwt, wip

LPR WER, F-score worse than HPR. # of words in LPR sets are much lower than HPR sets

ASER 2012 Decoding results

Improvements obtained on 2012 data

LM: 3 gram 2012 canonical ASER stories	**GM**: UP train+valid words (count>=2)						
Acoustic Model	Validation (UP+RJ) WER%	Validation F-score (P, R)	UP test WER%	UP test F-score (P, R)	RJ test WER%	RJ test F-score (P, R)	lmwt, wip
(VTLP warped + original) IITM Hindi Baseline; sp 1.1x 1.2x	23	0.96 (0.966, 0.954)	17.32	0.975 (0.976, 0.973)	19.63	0.97 (0.969, 0.97)	27, 1.0

ASER 2016 Decoding results (with/without denoising)

Improvements obtained on 2016 no story overlap with 2012 subset

LM: 3 gram 2016 canonical ASER stories	**GM**: UP train+valid words+ASER 2016 valid (count>=2)					
Acoustic Model	Undenoised WER% (valid,test)	Undenoised F-score (P, R) (valid, test)	lmwt, wip	DNS64 denoised WER% (valid, test)	DNS64 denoised F-score (P, R) (valid, test)	lmwt, wip
(VTLP warped + original) IITM Hindi Baseline; sp 1.1x 1.2x	**30.27** [5350 / 17677, 2372 ins, 412 del, 2566 sub] **34.50** [4376 / 12684, 1930 ins, 308 del, 2138 sub]	0.953 (0.965, 0.942) 0.949 (0.960, 0.938)	35, 1.0	**31.59** [5585 / 17677, 2584 ins, 381 del, 2620 sub] **36.38** [4614 / 12684, 2098 ins, 272 del, 2244 sub]	0.952 (0.967, 0.937) 0.947 (0.961, 0.933)	36, 1.0

2016 Data augmentation and chkwidth effects

More augmentations and reduced chunk-width: **Key contributors to improved performance**

Improvements obtained on 2016 no story overlap with 2012 subset

LM: 3 gram 2016 canonical ASER stories	**GM**: UP train+valid words+ ASER 2016 valid (count>=2)		
Acoustic Model	2016 WER% (valid,test)	F-score (P, R) (valid, test)	lmwt, wip
(VTLP warped + original) IITM Hindi Baseline; sp 1.1x 1.2x	**30.27** [5350 / 17677, 2372 ins, 412 del, 2566 sub] **34.50** [4376 / 12684, 1930 ins, 308 del, 2138 sub]	0.953 (0.965, 0.942) 0.949 (0.960, 0.938)	35, 1.0
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp) chkwidth=140	**24.70** [4367 / 17677, 505 ins, 1381 del, 2481 sub] **28.52** [3618 / 12684, 390 ins, 1160 del, 2068 sub]	0.926 (0.974, 0.883) 0.915 (0.972, 0.865)	38, -0.5

9-10% improvement in WER but minor improvements in F-score of correct words over baseline

ASER 2016 Decoding results

Improvements obtained on 2016 subset which has story overlap with 2012 set

LM: 3 gram 2012 canonical ASER stories   **GM**: UP train+valid words (count>=2)

Acoustic Model	WER%	F-score (P, R)	lmwt, wip from 2016 no story overlap experiment
(VTLP warped + original) IITM Hindi Baseline; sp 1.1x 1.2x	**35.90** [6772 / 18863, 2739 ins, 331 del, 3702 sub]	0.931 (0.968, 0.898)	35, 1.0
5(3)-0(8)-0.625(4)*1e-6; (2012)UP+RJ+hindi_CS (noise_aug+sp+pp+SpecAug) chkwidth=140	**26.62** [5022 / 18863, 1089 ins, 761 del, 3172 sub]	0.940 (0.967, 0.913)	31, 0.0

Improvements obtained with retraining. More insertions in baseline model, compared to more deletions in retrained model (seen throughout)

Discussion of Results

Going from 6 to 20 versions of data augmentation, minor improvements in 2012 data but large improvement (~2% in WER) in 2016 data because of speaker variations between regions.

Reducing chunk-width in 2016 no story overlap case leads to further improvements.

Reducing effect of text contexts during retraining helpful → No story overlap between 2012 retrain and 2016 test sets

Comparing the 2016 and 2012 sets, WERs and F-scores on 2016 sets worse:

Higher miscue rates and noisy nature of 2016 sets:

1) ASER test+valid 2016 data: 18 ON tags/min and 1.7 IR tags/min
2) ASER 2012 (UP+RJ test+valid) data: 14 ON tags/min and 0.8 IR tags/min.

Noises and irrelevant speech sections lead to a higher amount of insertions, substitutions and higher WER in the 2016 set compared to the 2012 set.

Discussion of Results

1. **Difference in WER between the baseline and the retrained model**:

More insertion counts in baseline; more deletions in retrained irrespective of lmwt, wip because of ON-SIL phone map during retraining. (IITM phone set has only one silence phone SIL)

Both models stumped a little by IR speech, retrained not so much by ON (other noise).

2. **2016 set**: Improvements in WER but only minor improvement in F-score with retrained model

Few correct words spoken in a sea of noise lost (decoded as SIL due to ON-SIL map)

Trade-off to be dealt with because of the ON-SIL mapping:

Better at ignoring noise in most cases for better WER but similar at detecting correct words in noisy 2016 recordings.

English Task & DAP lab English data

Summary of the DAP lab English datasets

Datasets	# of recordings	# of Unique stories	# of Unique speakers	Duration (min)
dahanu	622	18	149	653
cs-2016	821	35	23	262
cs-2019	830	36	81	467
df-ballaravada-2019	53	3	17	26
dfs-mumbai-2019	439	20	30	169
gbmc-2019	23	10	4	14
nashik-2018	100	6	15	44
shashwat-amravati-2019	67	1	33	62
st-michaels-ahmednagar-2019	60	1	30	50
vjhs-2018	181	18	43	87

English Task & DAP lab English data

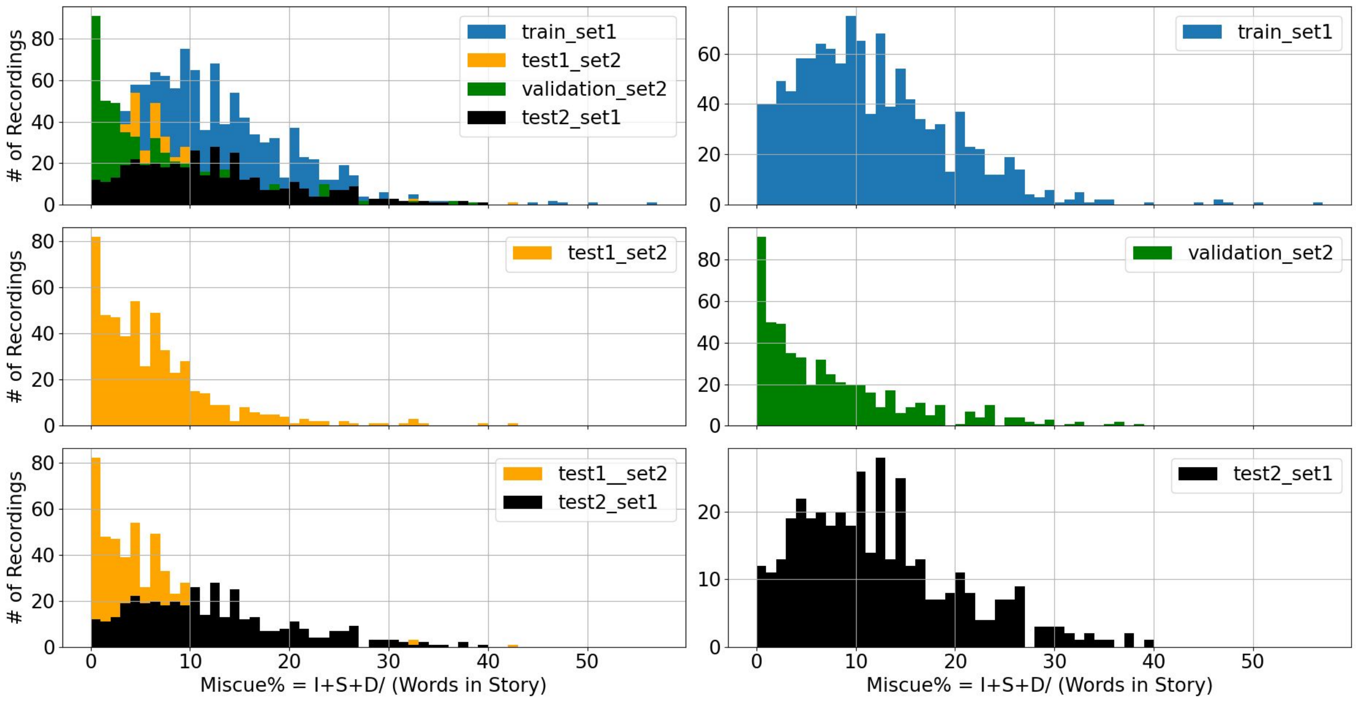
4: Summary of the DAP lab English data splits

Datasets	# of recordings	# of Unique stories	# of Non IR sentences	# of Unique speakers	Non IR Duration (min)
Train (Set 1)	1754	60	12869	278	1227
Valid (Set 2)	522	37	2689	45	193
Test1 (Set 2)	526	38	2868	56	203
Test2 (Set 1)	394	39	1925	46	189

Sentence split data used in all cases. Train and Test2 set have common stories (Set 1). Valid and Test1 have common stories (Set 2).

No story overlap otherwise and no speaker overlap as always between any of the splits

English splits miscue distribution



Test2 set has broader miscue range compared to test1 and valid

English decoding results

Improvements obtained on the English valid and test sets

LM: 3 gram All English canonical stories	**GM**: English train+valid words (count>=2)			
Acoustic Model	Validation WER% (No story overlap with DAP lab English Train)	Test1 WER% (No story overlap with DAP lab English Train)	Test2 WER% (Story overlap with DAP lab English Train)	lmwt, wip
(VTLP warped + original) IITM English Baseline; sp 1.1x 1.2x	**5.48** [1589 / 29022, 154 ins, 216 del, 1219 sub]	**5.52** [1676 / 30355, 220 ins, 247 del, 1209 sub]	**12.94** [2902 / 22422, 260 ins, 506 del, 2136 sub]	23, 0.0

Discussion of results

Improvements obtained only after reducing chunk-width because reduced effect of the text contexts during retraining particularly helpful:

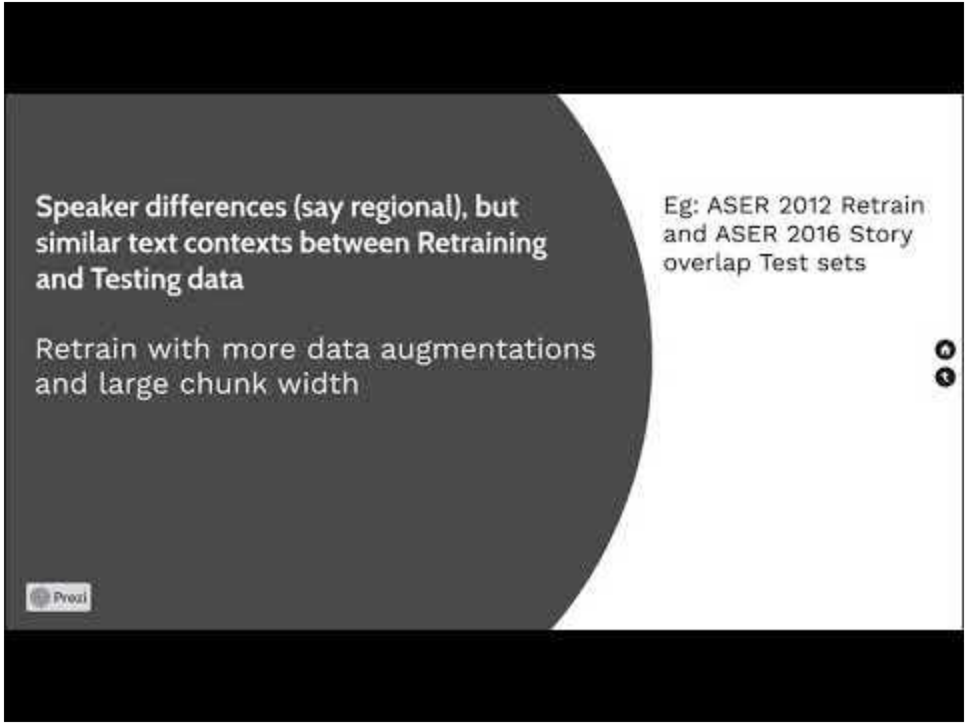
1) No story overlap. Similar to ASER 2016 No story overlap scenario
2) Larger variety of stories (8-12 in ASER vs 30-40 English stories)
3) Reduced size of speaker set compared to ASER Hindi (1000 in ASER vs 250 in English)

In the 140 chkwidth retrained model, drop-off in WER on Test2 set (which has story overlap with retrain set) is the lowest: 0.1% compared to 1-1.5% on valid and Test1 set

The reduced chunk-width is a way to eliminate the inherent text contexts present in this retraining data and “clean” it for a general transfer learning purpose. It controls a new aspect of transfer learning and changes the construction of the mini-batch itself.

Summarizing transfer learning scenarios

Video/Animation that summarizes scenarios examined



Comparison with off-the-shelf ASR systems

Results from a recent system testing by Pratham on a 100 paragraph test set (ASER 2012) :

Total # of reference words: 4755

Speech to text (STT) system	Google	Azure	Azure CT	**IITB ASR**
WER (%) (scored with sclite)	28.3	23.1	24.1	13.0
Precision (%)	99.02	98.76	98.34	97.02
Recall (%)	83.30	89.45	90.05	98.98
F-score (%)	90.48	93.87	94.01	97.99

The STT output is post-processed for alignment with the canonical text in order to obtain the words uttered correctly.

Conclusion and Future Work

Techniques of data augmentation and transfer learning were investigated

Improvements in miscue detection obtained on test sets over a baseline model

Data augmentation useful when speaker dissimilarities are prominent

Reducing chunk-width for curtailing effect of text context particularly helpful

FUTURE WORK:

Train a (simpler) denoiser using ASER samples for better noise profile

Cross language transfer learning (from Hindi to Marathi). Initial experiments: mixed results, More experiments on English transfer learning

Further improvements that can be made to the LM (sub-word modeling) and GM.

Submissions made to IITM ASR challenges

IITM ASR Challenge	Test set WER (%)	Approach	Ranking
Hindi closed task	7.47	kaldi TDNN chain model + RNNLM	7
Hindi open task	9.48	“” + fine tuned on dev	5
English closed task	5.33	kaldi TDNN chain model + RNNLM	2
English open task	5.27	“” + fine tuned on dev	3

References

[1] ASER. ASER Report for the year 2018. http://img.asercentre.org/docs/ASER2018/ReleaseMaterial/aser2018nationalfindingsppt.pdf, Last accessed October 2020

[2] The Hindu. Basic literacy, numeracy skills of rural Class VIII students in decline.https://www.thehindu.com/news/national/basic-literacy-numeracy-skills-of-rural-class-viii-students- on-a-decline-aser-2018/article26004114.ece, Last accessed October 2020

[3] Speech Processing Lab IIT Madras. Hindi ASR challenge. https://sites.google.com/view/asr-challenge, last accessed: September 2020

[4] Dolly Agarwal, Jayant Gupchup, and Nishant Baghel. A dataset for measuring reading levels in india at scale. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 9210–9214. IEEE, 2020

[5] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, volume 117, 2013

References

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779, 2019

[7] Xingchen Song, Guangsen Wang, Zhiyong Wu, Yiheng Huang, Dan Su, Dong Yu, and Helen Meng. Speech- xlnet: Unsupervised acoustic model pretraining for self-attention networks. arXiv preprint arXiv:1910.10387, 2019

[8] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association, 2015

[9] Kevin McGuinness. 2nd Workshop on Deep Learning for Multimedia, Insight Dublin City University . https://github.com/telecombcn-dl/2018-dlmm/raw/master/D2L02 Transfer.pdf, Last accessed October 2020

[10] Prashanth Gurunath Shivakumar and Panayiotis Georgiou. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. Computer speech & language, 63:101077, 2020.

Thank you

Extra slides: Examples of decoded texts and discussing results

1. **biju_aser_cg_S3-P_2 (Example for more insertions in baseline)**

(a) Ground Truth: **SIL ON IR ON आज मामा आए**

(b) Baseline model decoded text: **हँ स चा बड़ी पर लौट रंग की गया आज मामा आए**

(c) Retrained model decoded text: **पढ़ने लौट है आज मामा आए**

Fewer words in the retrained model text. When evaluating WER: GT is just ”आज मामा आए “ so fewer insertions in retrained model text

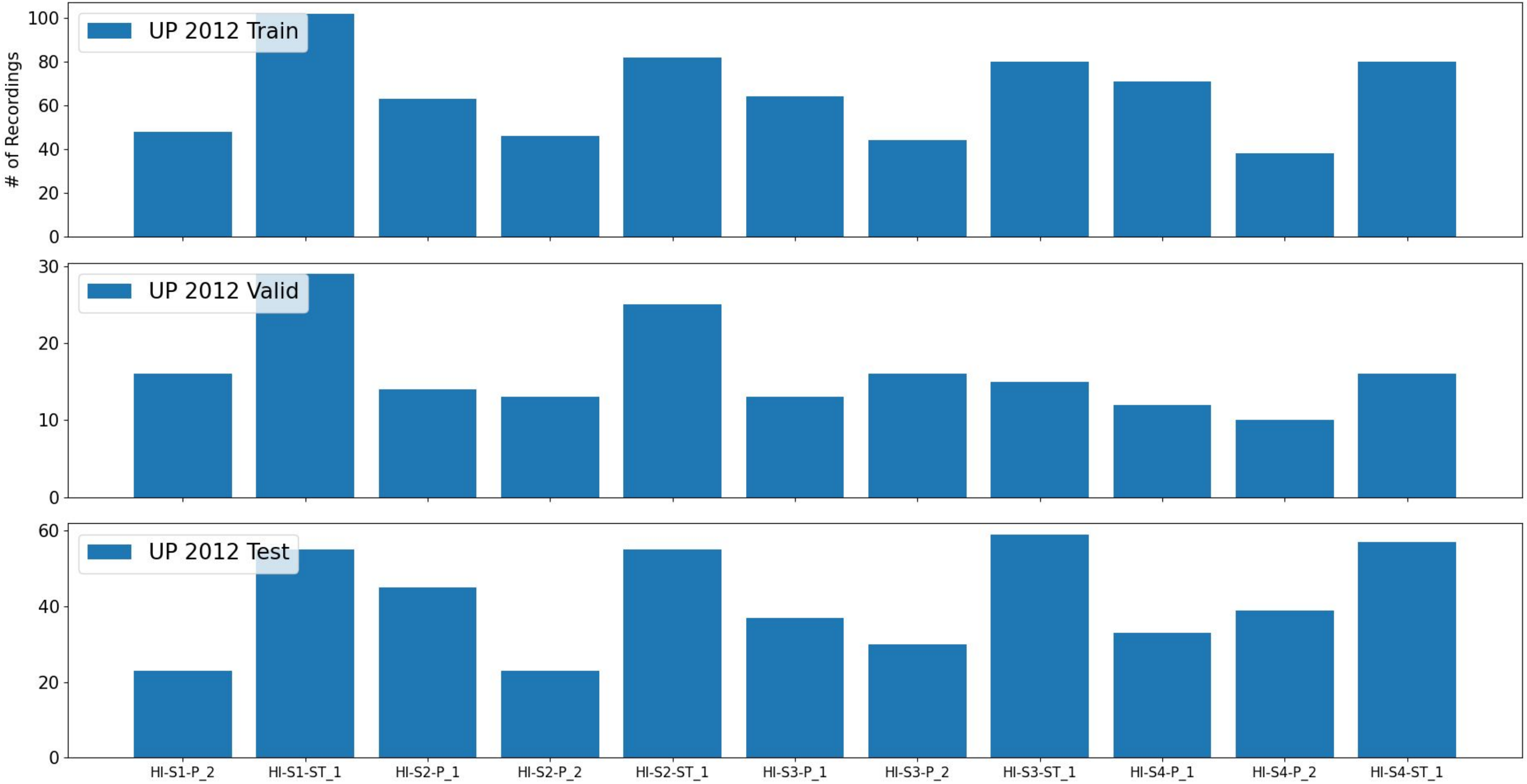
2. **mukesh_aser_uk_S4-P_2 (Worst case example for only slight improvement in F-score)**

(a) Ground Truth: **ON IR ON मोर ON मोर चाचा की MB ON सादी हुई ON IR ON सबको ON नई ON ON IR FP ON IR**

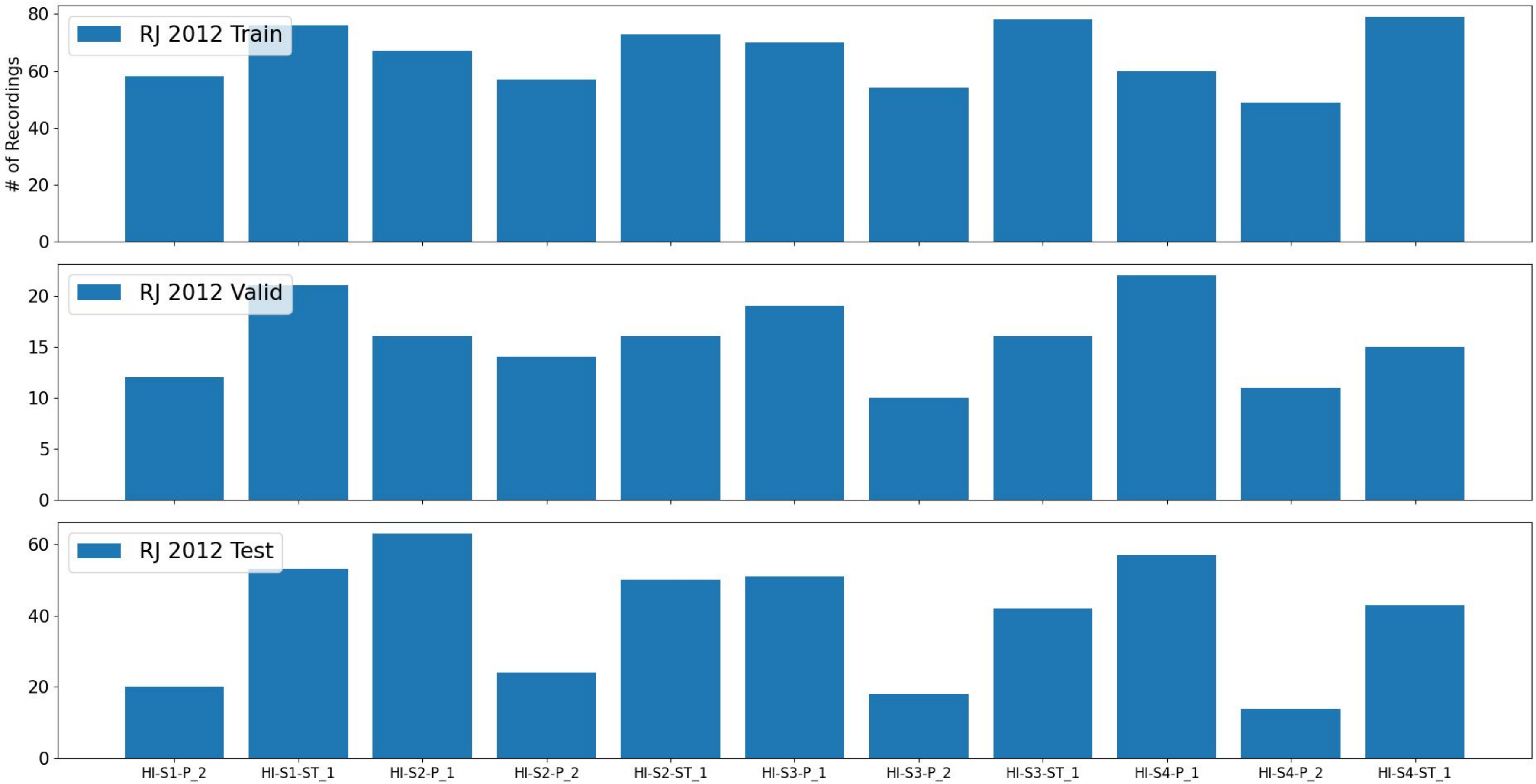
(b) Baseline model decoded text: **हर साथ्यों मौज चाँ द सुरन की में एक लगाए हुई आती ही मोर खाकर रही थी सब को नी मं गाकर यह गाय ह**

(c) Retrained model decoded text: **की सब को नहीं**

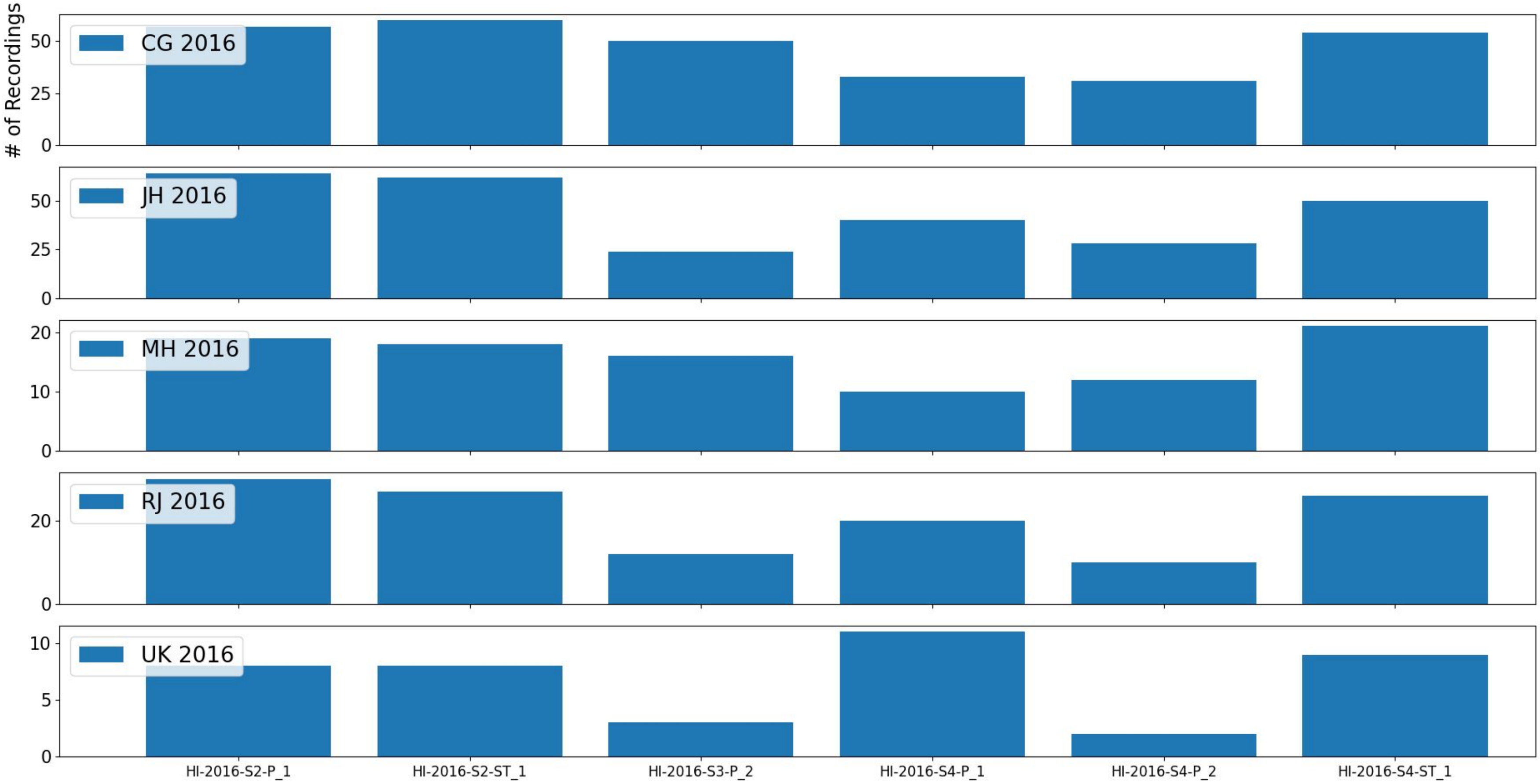
Extra slides on story distribution



Extra slides on story distribution



Extra slides on story distribution



Extra slides on story distribution