Task-specific Language Modeling for Oral Reading Assessment

Kanhaiya Kumar
13D070046
Guide: Prof. Preeti Rao

Introduction

- Nearly 70% of India's population lives in rural areas. Literacy skills among children are very poor.
  - Majority of Std. V students cannot even read Std. II level of text[1].

- Problem source: Huge shortage of skilled teachers.

- Need a scalable technological solution which facilitates Oral reading practice & assessment

[1] ASER: The Annual Status of Education Report (rural). http://img.asercentre. org/docs /Publications/ASER%20Reports/ASER_2012/fullaser2012report.pdf. ASER Centre 2012

Introduction

- Overall Goal:
  - Design an efficient and robust automatic assessment system for the reading ability.

- This automatic assessment can be categorizes into two parts:
  - Word-level assessment
    - Detecting Word-level miscues
  - Speech delivery assessment
    - Measuring speech rate, fluency, prosody

Block Diagram

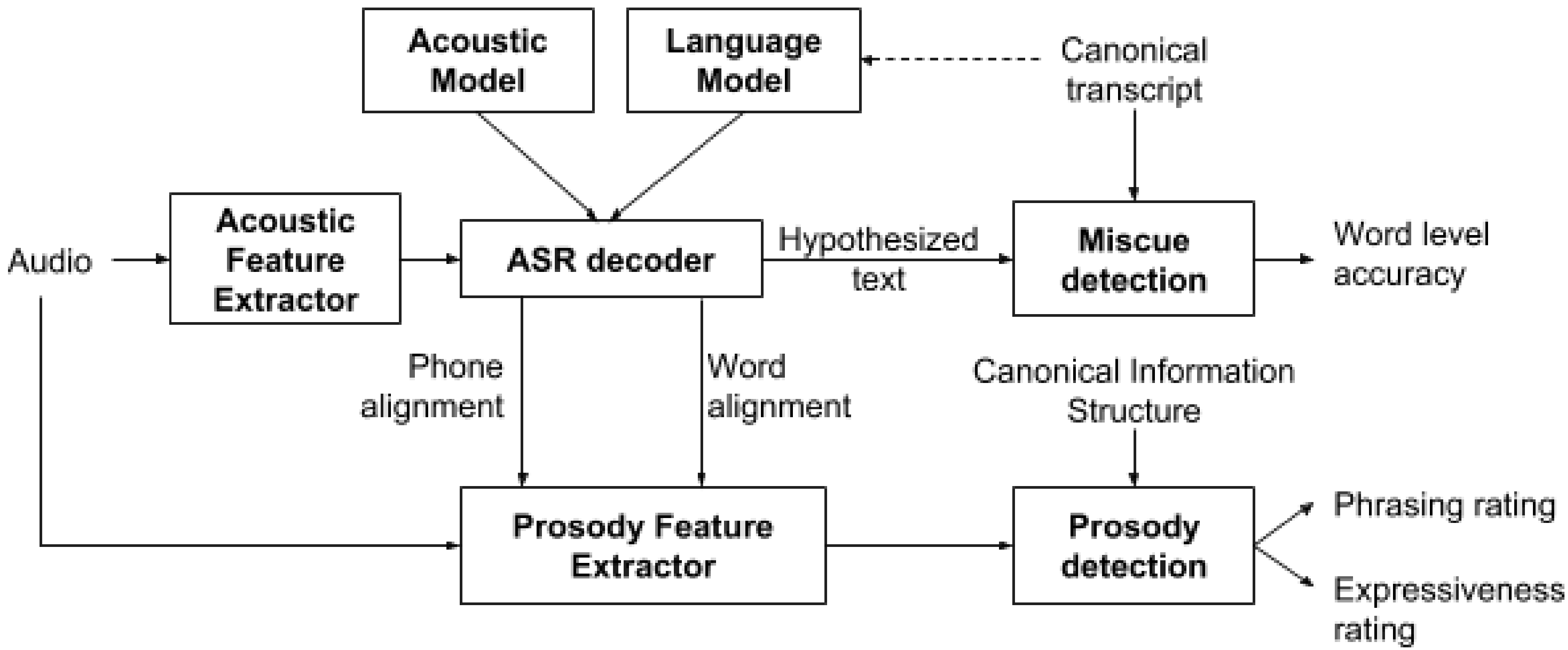


Fig: Overall System block diagram

Assessment Results

Story Name: One Good Turn Deserves Another

Speaker Name: TDC

Speaker Audio Model Audio

Lexical Evaluation

one good turn (one) deserves (good) another (crowed) (this) (our) (and) (an) (other) fred

was a farm worker (were) (a) who found a young (used) eagle caught (yak) in a trap (flapping)

(ends) (trick) he couldn't (he) bear (could) (not) (where) to see such a beautiful bird

in pain (plan) (a) so he released (realized) it a few days later he was sitting (still)

in (a) (in) the shade of an old wall having bread (a) (bird) and cheese for his lunch

Total miscues: 28

C S I D

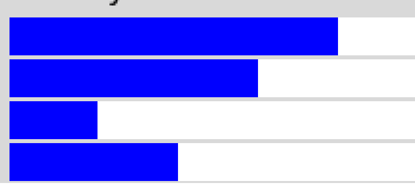
Prosody Evaluation

Speed:

Phrasing:

Expressiveness:

Meaning:



How an ASR system works

What is the most likely Word sequence given acoustic observations O?

$$\hat{W} = \arg\max_{W} P(W|O)$$

$$\hat{W} = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)}$$

$$\hat{W} = \arg\max_{W} P(O|W)P(W)$$

Acoustic Model

Language Model

Datasets

- Speech data read by students from our campus school, of age group 10-14 years
- Training speech data:
  - Used in training acoustic model
  - 57 Hindi and English stories read by 41 fluent English and Hindi speakers
  - comprising 5.2 hours of speech data
- Training text data:
  - Used in LM model training
  - 80 Hindi and English stories text
- Evaluation data:
  - 15 English stories read by 3 dis-fluent speakers
  - comprising 30 utterances each of ~1 min duration

System Evaluation Metrics

- Problem with Word Error Rate (WER)
  - “hunter” → ”hunters”, will be considered as a substitution error.
- Phone Error Rate (PER)
  - Compares the similarity of strings at the phone level
  - Converted all the words into phone sequence using word to phone mapping dictionary
- Miscue Detection:
  - 3 types of miscues: Substitution, Insertion and deletion
  - Backtracking path from edit-graph will give a CSID sequence

System Evaluation Metrics

- Canonical text : We were very happy
- Ground truth text: We where a very happy happy
- Hypothesized text: We were aware happy happily

	"CSID" sequence	Miscue & Non-Miscue sequence
Ground-truth	CSICCI	CMCM
Hypothesized	CCSCI	CCMM

- 1TP, 1 TN 1FN and 1 FP
- Both the miscue detection and false alarm rate will be 50% in this case.

Performance of Google Speech Engine

We can provide the context of the audio to the Google speech API

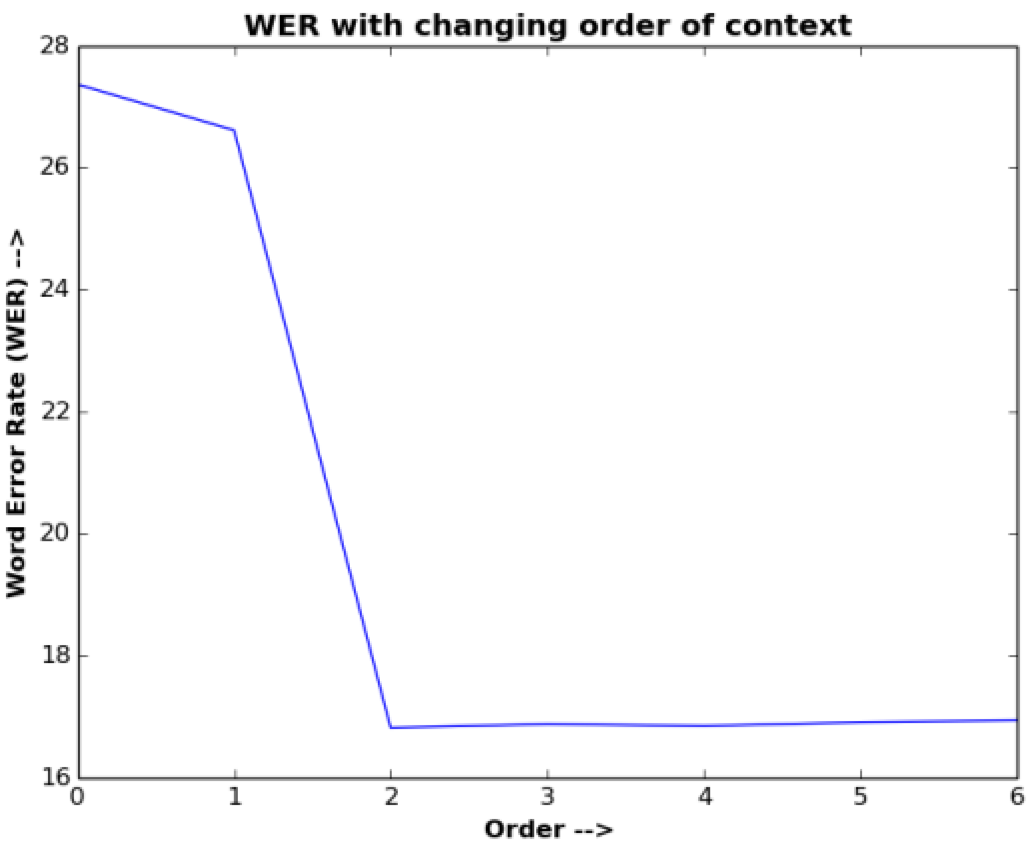


Fig: WER on Evaluation data at different order of context

- ❖ For order 1:
  - ➢ DR=59.34% at 15.81% FAR
- ❖ For order 2:
  - ➢ DR=43.40% at 4.10% FAR

Challenges with using Google Speech Engine

- Minimum WER achieved is 16.81% still very high
  - Getting 12.12% with canonical only!
- Can not get phone level alignment
  - it gives word level alignment but that itself is not correct (includes the silences)
- Getting only 43.4% miscue detection rate at 4.1% false alarm rate
  - Very low FAR required specially for this task
- Continuous Internet connection
  - Difficult for rural areas in India
- It’s a paid system

Acoustic Model

- Can use the conventional AM as used in general ASR system
  - Because task-specific constrains is only for LM

- Used Deep Neural Network(DNN) based acoustic model for our task
- fMLLR transformed features are inputs to this DNN which gives probability of each phone
- Variant of this probability is being used as the emission probability in the decoding graph.
- The AM is developed in previous work[2] using the Kaldi framework

[2] P. Swarup. "Acoustic model training and adaptation for children's read speech recognition". M.Tech dissertation, Department of Electrical Engineering, IIT Bombay, 2017.

Guided Language Modelling in Literature

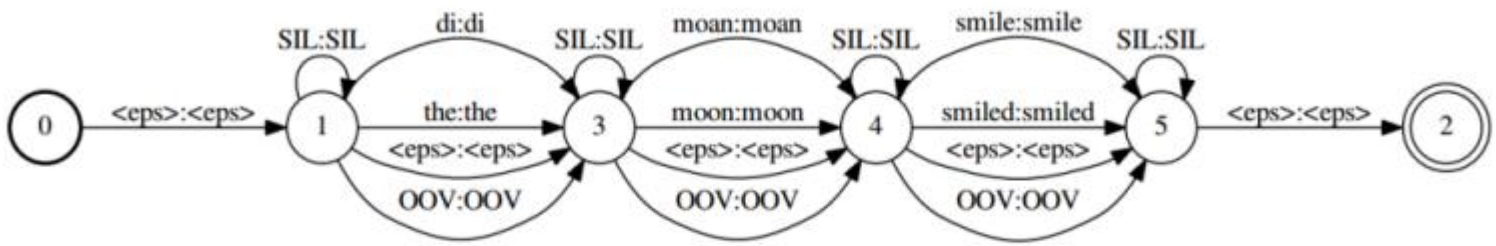


Fig: LM for the sentence ”The moon smiled” [3]

Problems:

- Finding Expected substitution could be an exhaustive task
- No back-loops (no repetitions path)
- Inhaling problem of OOV
- Segmentation into sentences required, any error will add into the recognition accuracy

[3] P. Swarup H. Tulsiani and P. Rao. ”acoustic and language modeling for children’s read speech assessment”. Proceedings of National Conference on Communications, Chennai, India, 2017.

Guided Language Modelling in Literature

- Target (trigram) model trained on current story
- Garbage(unigram) model trained on general domain text
- Built using Context Free Grammar(CFG)

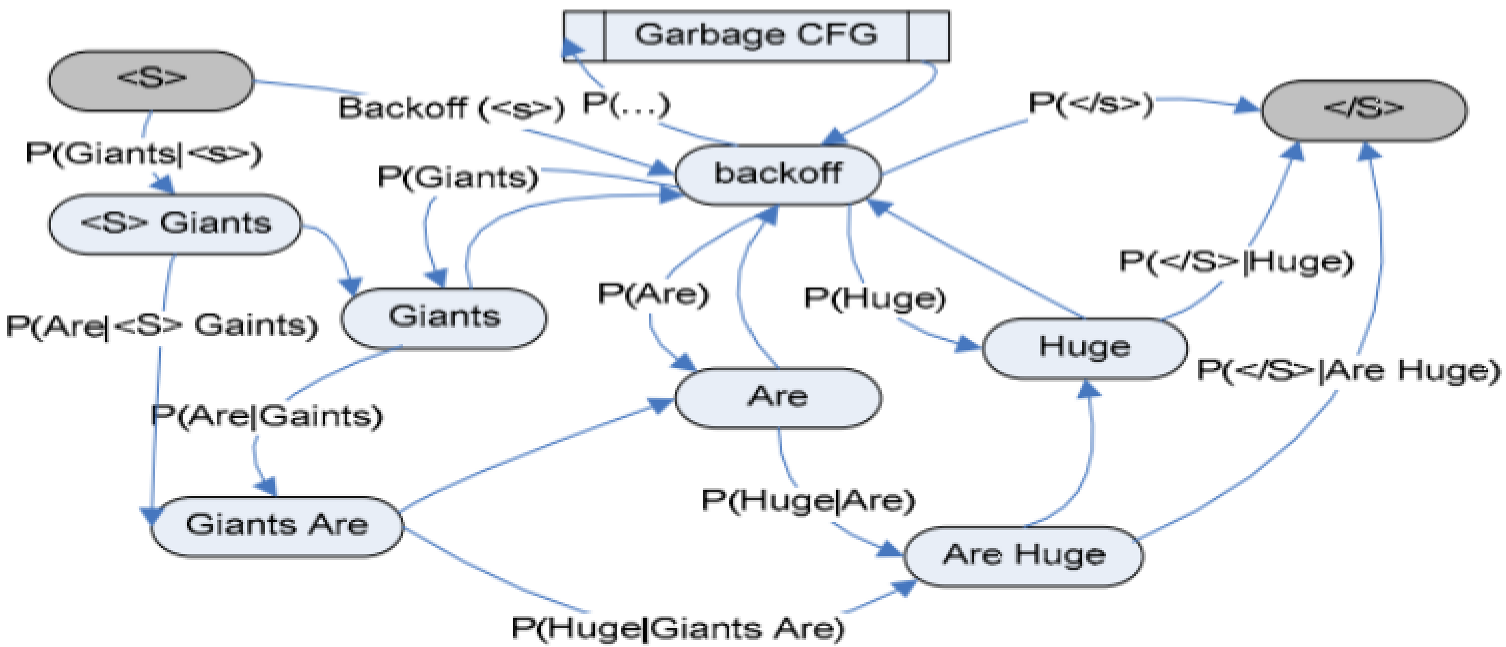


Fig: FST of Trigram model for the sentence ”Giants are huge” along with garbage model[4]

[4] Yun-Cheng Ju Xiaolong Li, Li Deng and Alex Acero. “Automatic children’s reading tutor on handheld devices”. Interspeech 2008.

Proposed Language Model

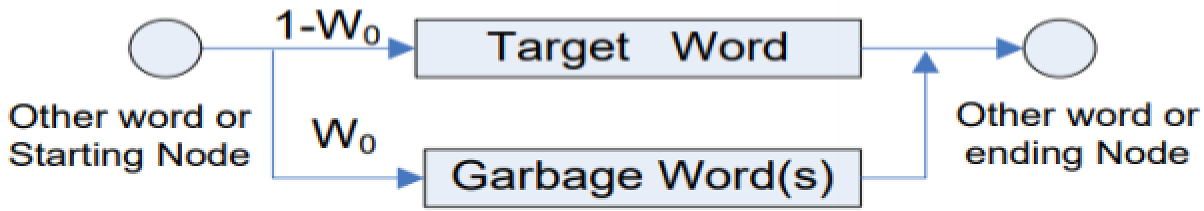


Fig: Target & Garbage based model[4]

Our proposed LM architecture

- We have used "zero-gram" LM in garbage model i.e. giving all words equal probability

Why?

- Mispronounced word by a child may not follow a unigram model.
- e.g. "jumped" could be pronounced as "jump"+"aid",
  - Here "aid" is not as frequent as the word "the" (why should we give higher probability to "the")
  - The least we could do is to assign equal probability to all

[4] Yun-Cheng Ju Xiaolong Li, Li Deng and Alex Acero. "Automatic children's reading tutor on handheld devices". Interspeech 2008.

N-Gram Language Models

- For the Word Sequence: $W = w_1, w_2, w_3, ...w_n$

$$P(W) = p(w_1, w_2, w_3, ..., w_n) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)...p(w_n|w_1, w_2, ...w_{n-1})$$

Using Markov assumption,

- For Bigram Language Model

$$p(w_1, w_2, w_3, ..., w_n) \approx p(w_1)p(w_2|w_1)p(w_3|w_2)...p(w_n|w_{n-1})$$

$$p(w_2|w_1) = \frac{Count(w_1, w_2)}{Count(w_1)}$$

What if we have a zero count?

N-Gram Language Models

Interpolation:

- Weighted interpolation of trigram, bigram and unigram counts

$$P_I(w_n, w_{n-1}, w_{n-2}) = \lambda_1 P(w_n, w_{n-1}, w_{n-2}) + \lambda_2 P(w_n, w_{n-1}) + \lambda_3 P(w_n)$$

Back-off:

- We will back-off to the lower order N-gram only if we have zero counts of the current N-gram

$$P_B(w_n, w_{n-1}, w_{n-2}) = \begin{cases} \tau(w_n, w_{n-1}, w_{n-2}) & \text{if} \quad count(w_n, w_{n-1}, w_{n-2}) > 0 \\ \gamma(w_{n-1}, w_{n-2}) P_B(w_n, w_{n-1}) & \text{if} \quad count(w_n, w_{n-1}, w_{n-2}) = 0 \end{cases}$$

N-Gram Language Models

- These extra probabilities assigned to the unseen n-gram will disturb the overall probability sum
- Discounting factor are usually introduced within each n-gram to compensate for the overall probability sum.

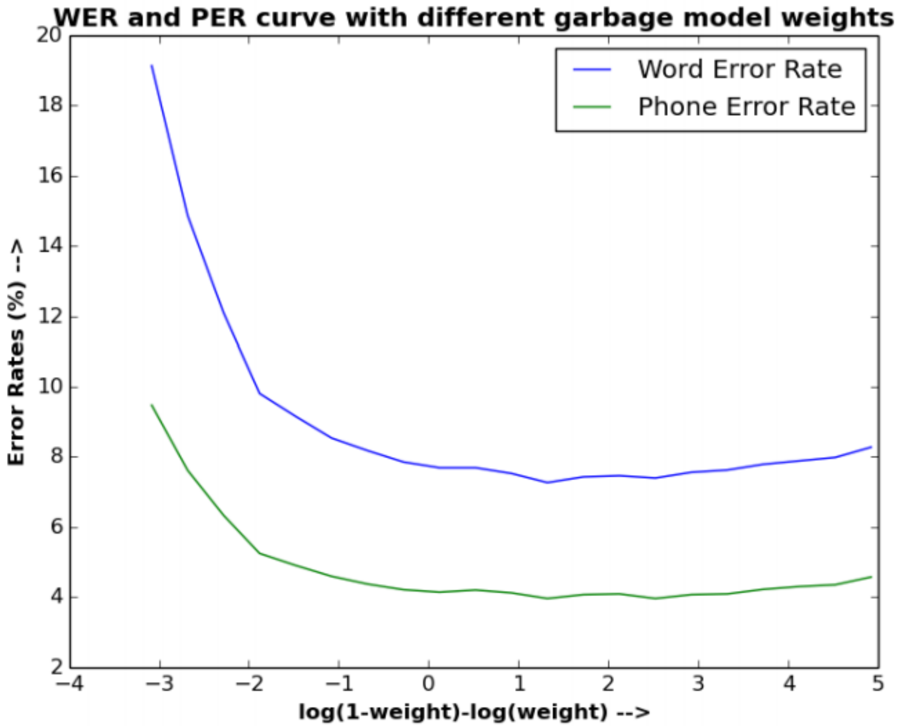
This can be done in two different ways:

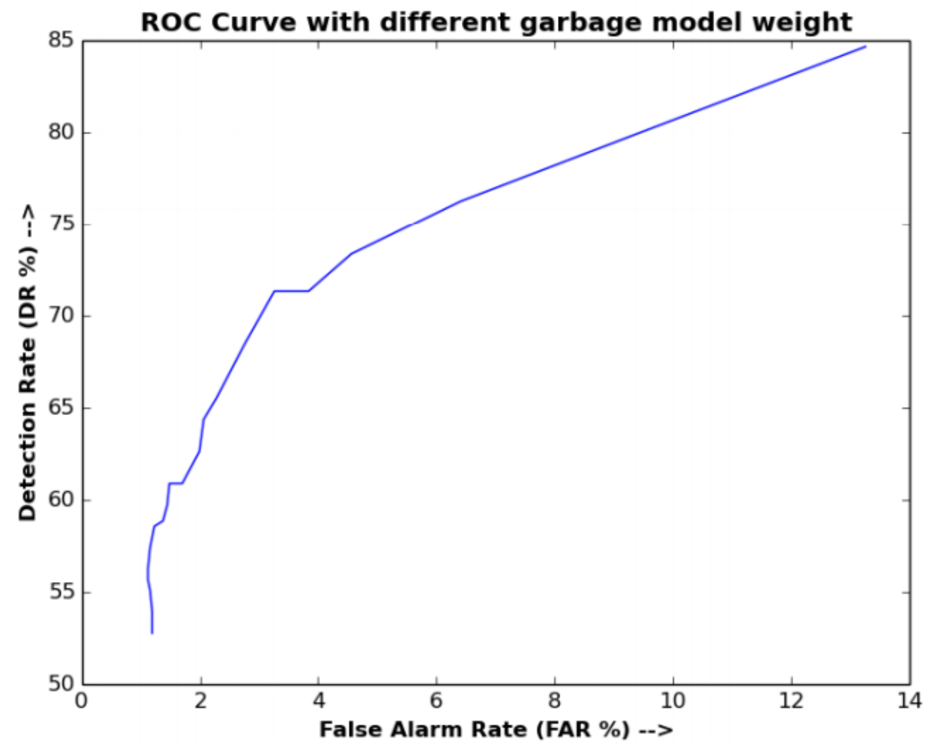
- Improved Kneser-Ney Smoothing:
  - Discounting is done by subtracting from the numerator
- Witten-Bell Discount:
  - Discounting done by adding into the denominator

Toolkits Used

- IRSTLM Tool
  - Used to get the N-gram probabilities given a text file
  - Can change order or smoothing methods
- Openfst Tool
  - Used to build the FST corresponding to the above N-gram probabilities
- Kaldi scripts
  - Used to make the overall decoding graph using the above LM
  - Used to build acoustic model
  - For decoding on the graph

Results at different garbage model weight

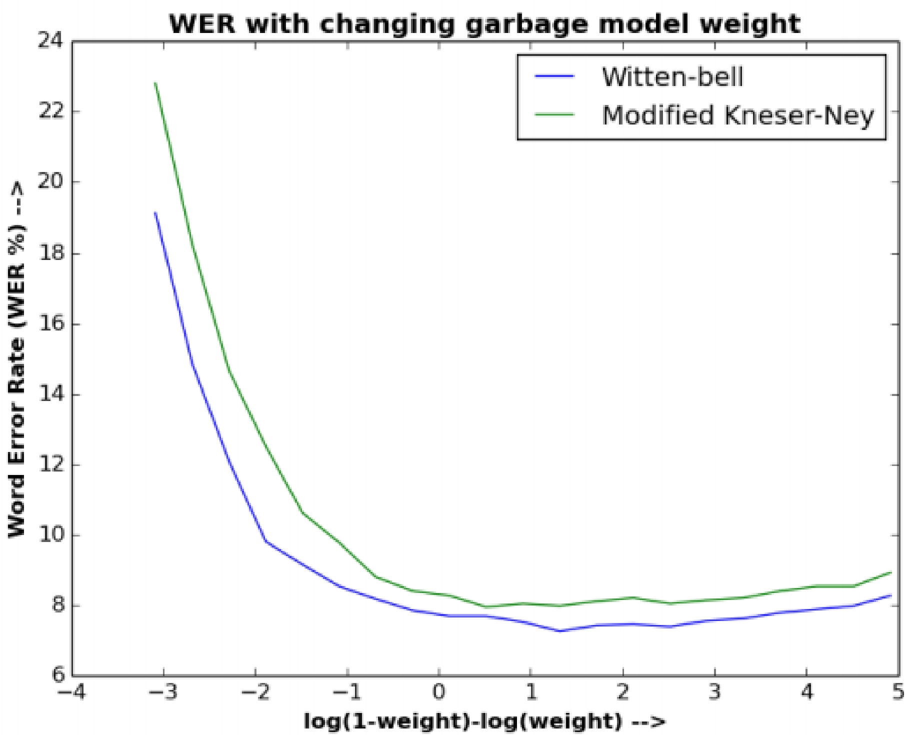


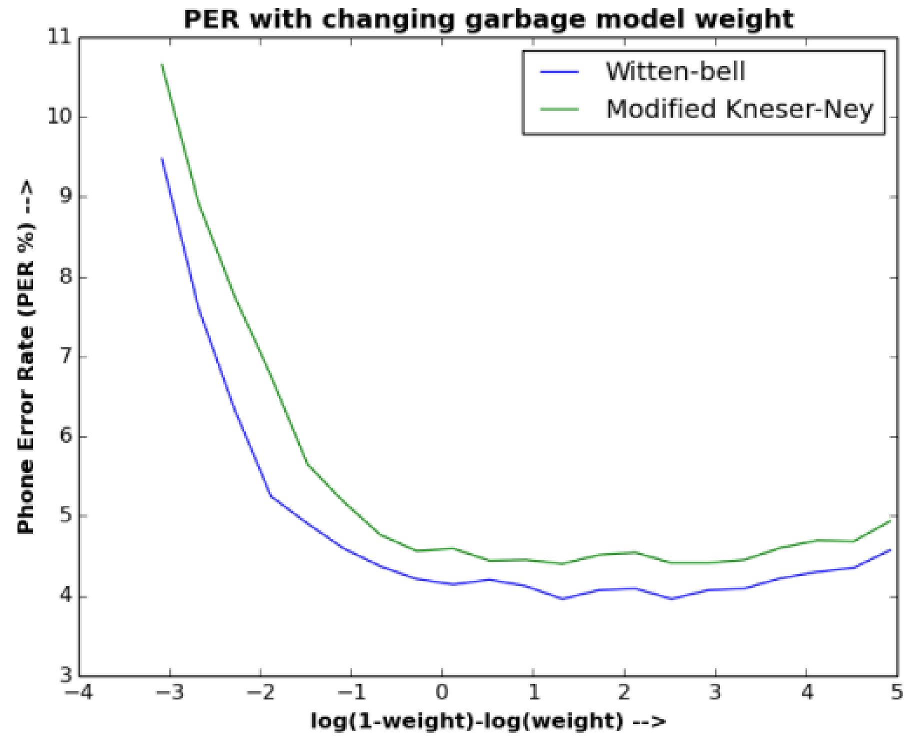


- The LM model uses:
  - Target model : Trigram on current story
  - Garbage model : zero-gram on 3000 words

- Getting minimum WER & PER as 7.26% and 3.95% resp.

Results with different smoothing algorithms



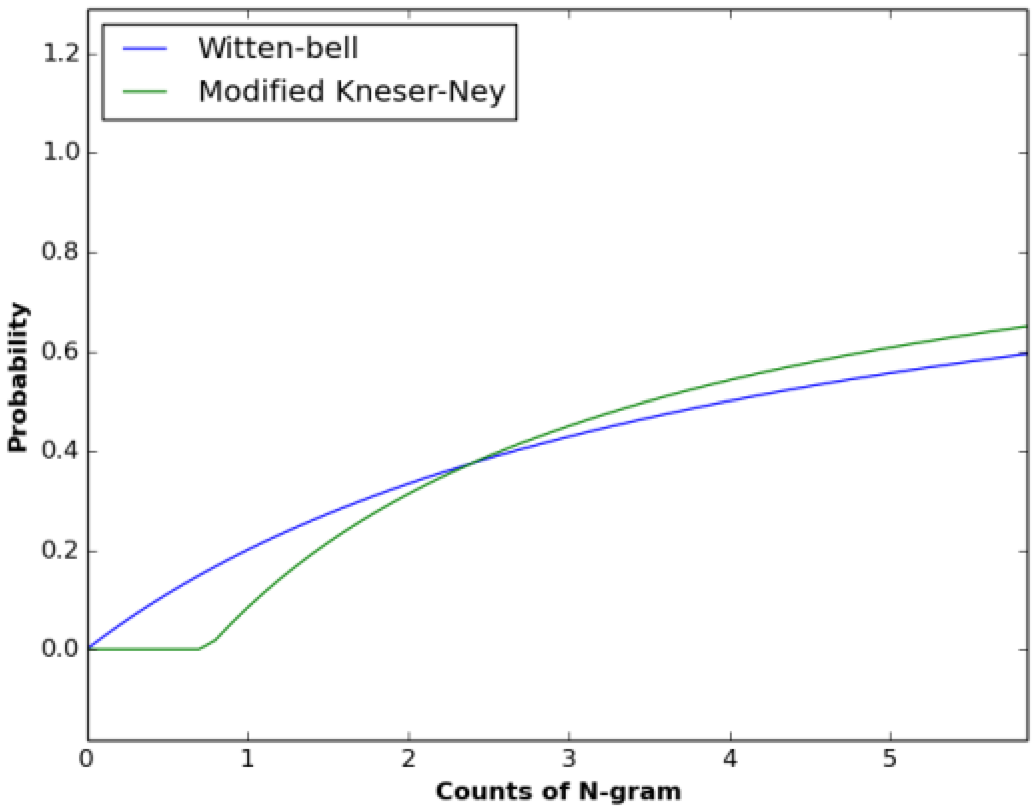


- ❖ Witten-bell performs better than Modified Kneser-Ney!
- ❖ However, [5] shows Modified Kneser-Ney performs better using perplexity

[5] Ismail. "Comparison of Modified Kneser-Ney and Witten-Bell Smoothing Techniques in Statistical Language Model of Bahasa Indonesia", 2nd International Conference on Information and Communication Technology (ICoICT), May, 2014.

Why Witten-bell performs better here?

$$y = \begin{cases} \frac{x}{N+x+C} + B & \text{for Witten-Bell} \\ \frac{max(x-d,0)}{x+C} + B & \text{for Modified Kneser-Ney} \end{cases}$$

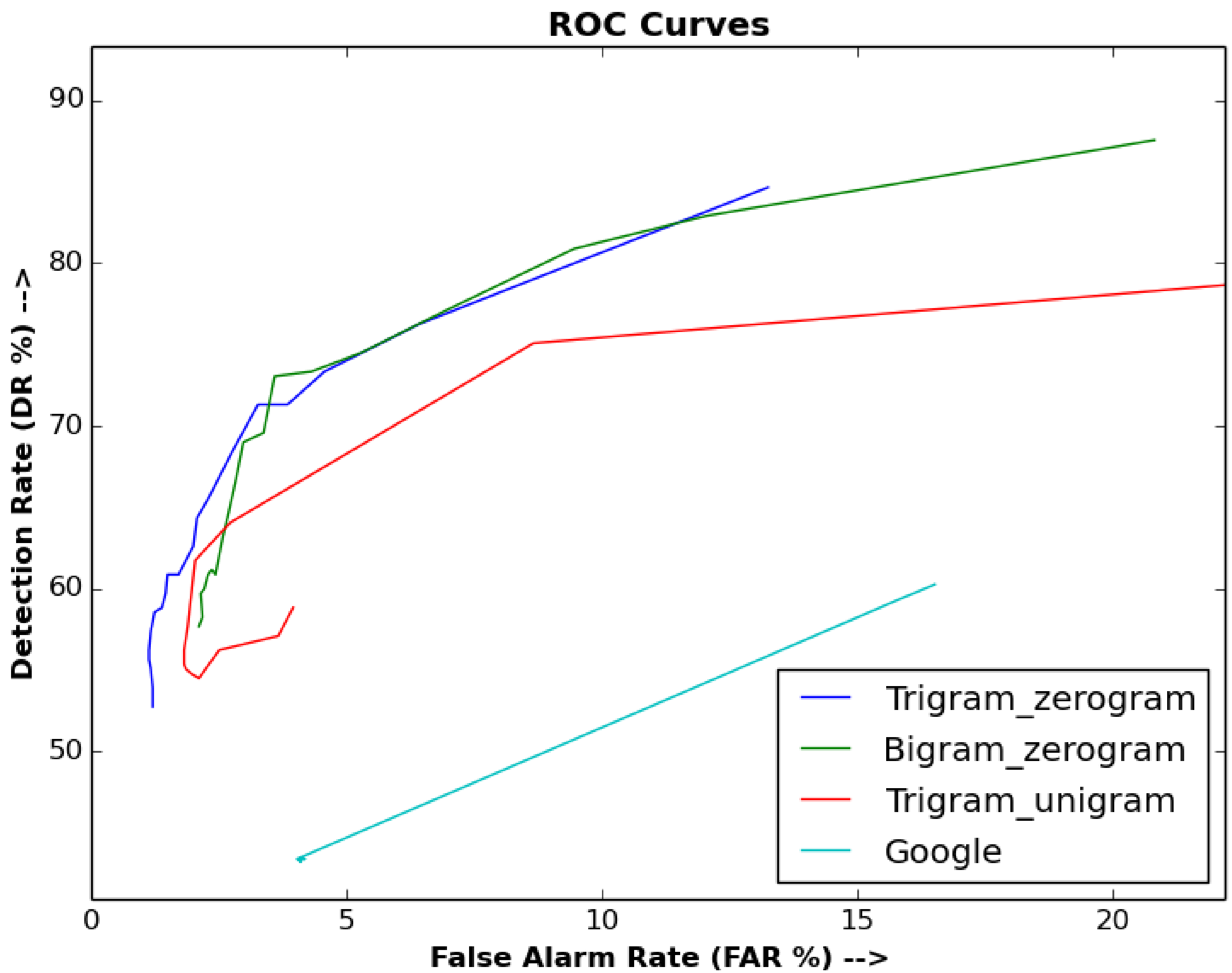


- Witten-Bell gives relatively high probability to lower counts than Modified Kneser-Ney

Different LM architectures

- Trigram zero-gram
  - Target model is tri-gram and garbage model is zero-gram
- Bigram zero-gram
  - Target model is changed to bi-gram while garbage model is still zero-gram
- Trigram unigram
  - Target model is tri-gram but the garbage model is uni-gram.
- All trigram
  - Here the LM is just a trigram model trained on 80 Hindi and English stories(No garbage model has been used)
- All bigram
  - Similar to All trigram model, here we have used bigram model to train.

Results with different LM architectures

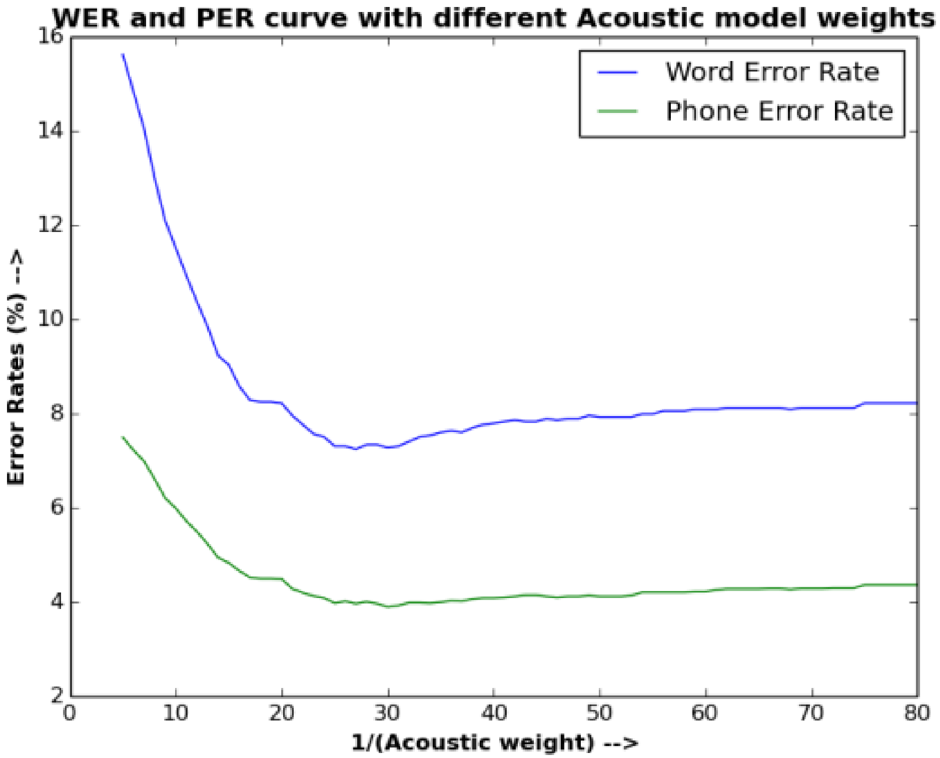


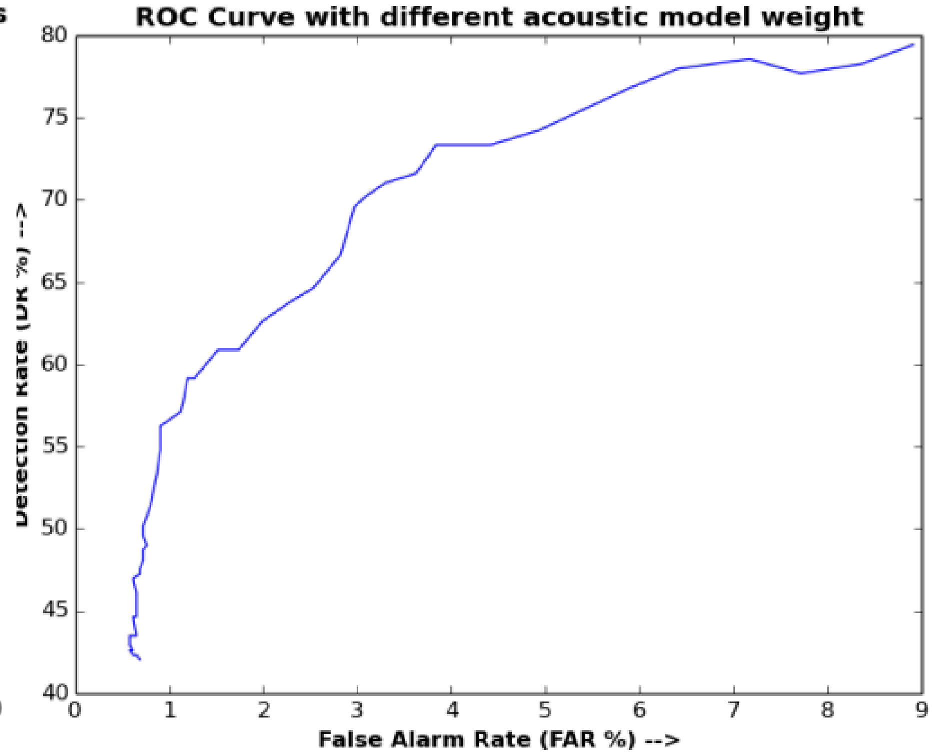
Results with different LM architectures

Sr. No.	Target N-gram	Garbage N-gram	Smoothing Algorithm	WER (%)	PER (%)	DR(%) at 5% FAR
1	Trigram	zerogram	Witten-bell	**7.26**	**3.95**	74.03
2	Trigram	zerogram	Modified Kneser-Ney	7.95	4.44	73.73
3	Trigram	unigram	Witten-bell	8.08	4.25	68.47
4	Bigram	zerogram	Witten-bell	8.08	4.42	**74.15**
5	All Trigram		Witten-bell	8.60	4.70	(63.18, 2.42)
6	All Bigram		Witten-bell	10.55	5.87	(71.59, 5.24)
7	Google's with bigram context			16.81	-	(43.4, 4.1)

Table: WER , PER and miscue DR at 5% FAR for different LM architectures

Results with different AM weights

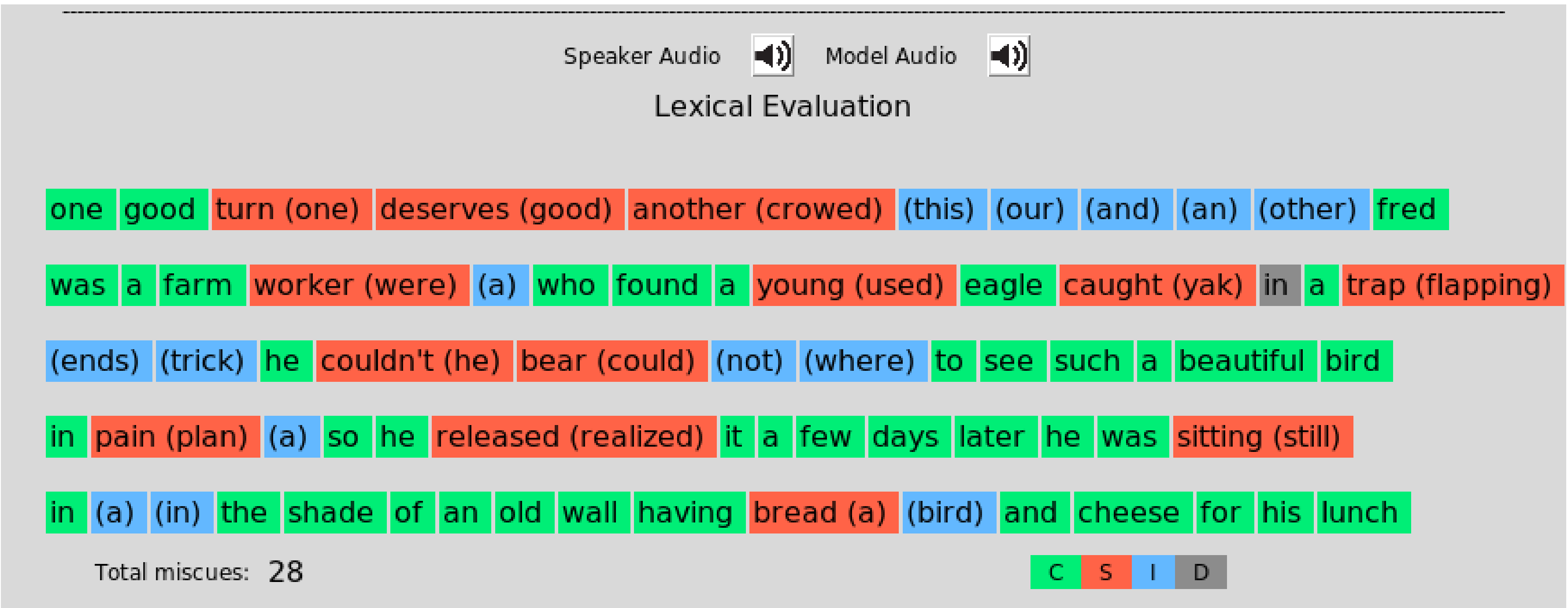




- Keeping the LM fixed at best model (i.e. Trigram target and zero-gram garbage model)
- The minimum WER & PER achieved at 1/27 acoustic weight
- Gives 7.23% and 3.95% WER and PER resp.
- Miscue detection rate also got increased to 74.42% at 5% FAR

Results with different quality of utterance

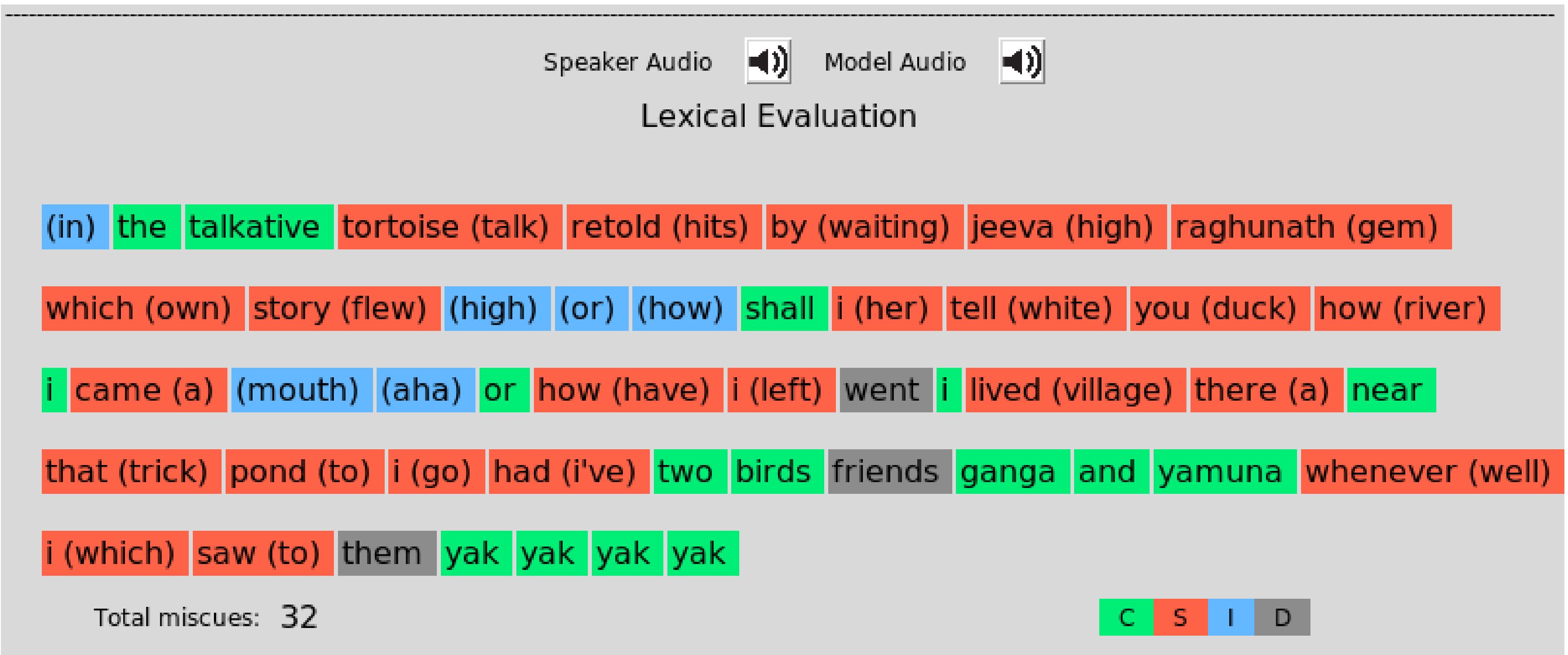
- Hesitations(repetitions), Wrong words



- Performing very well for urban accent
- The wrong words are getting substituted with similar sounding words (specially “realized” in middle of 4th line)

Results with different quality of utterance

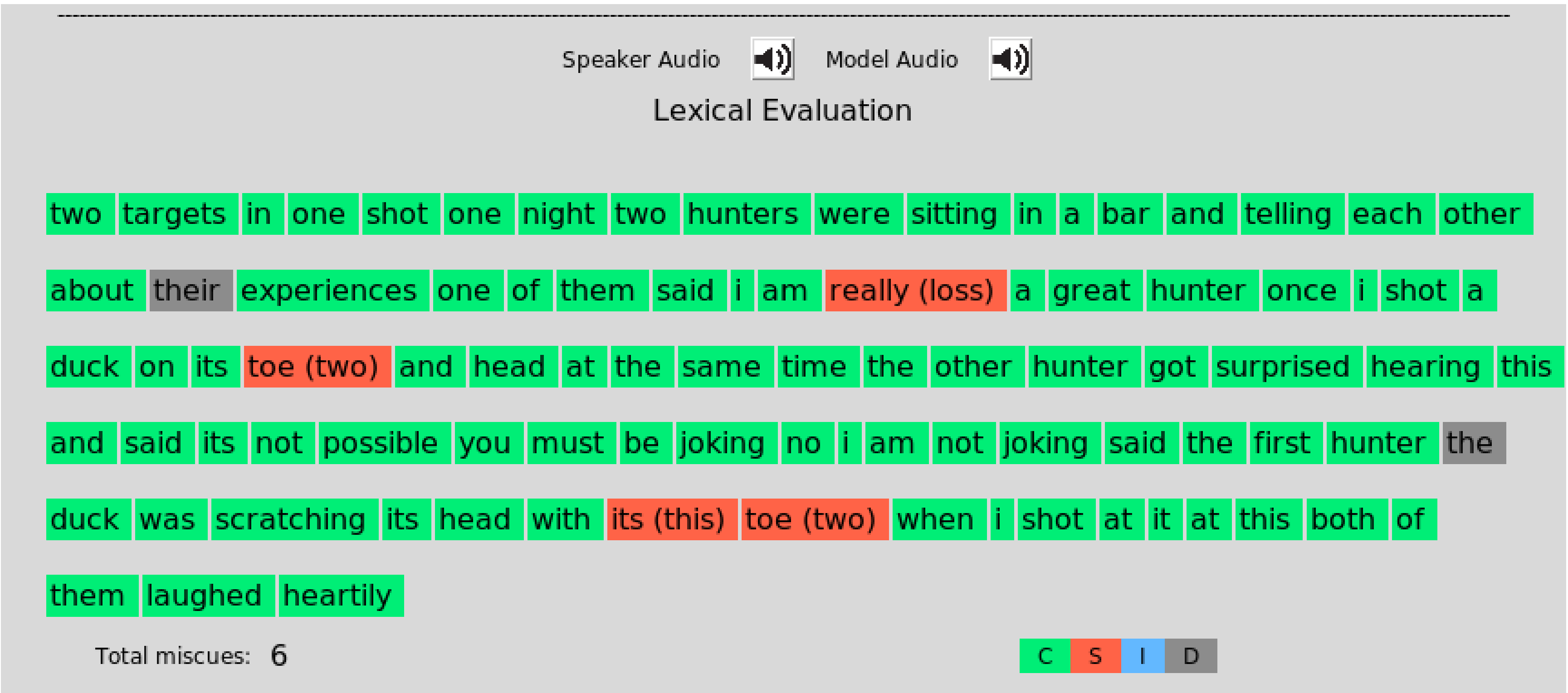
- Sound-outs, Long pauses, Hesitations(repetitions of partial words)



- Predicting small words whenever there is a sound out
- But, Correct pronunciations are getting recognized correctly
- Long pauses are correctly handled (in 4th line, “i’ve two birds”)

Results with different quality of utterance

- Fluent, minor skips, High speech delivery rate



- For very fast speakers, it is making few mistakes (skipped “their” in the 2nd line)
- “toe” has been replaced with “two”,
- When we reduce the Acoustic weight, “toe” was getting decoded correctly

Conclusion

- We proposed a task-specific Language model for the children reading assessment task.
- On our data, the proposed model gives 7.26% WER and 3.95% PER, and we are getting around 74.42% miscue detection rate at only 5% FAR.
- No need of any task-specific or story specific annotated data as opposed to the previous work[3] .
- Can add any words (even the Hindi words) in the garbage model without the training text
  - As opposed to the uni-gram garbage model, where the words are constrained by the training data

[3] P. Swarup H. Tulsiani and P. Rao. ”acoustic and language modeling for children’s read speech assessment”. Proceedings of National Conference on Communications, Chennai, India, 2017.

Conclusion

- Generalizable because of very different training and Evaluation data
  - The acoustic and garbage weight can be learned on small development set
- Advantage of setting the hyper-parameter for desired FAR

Future Works

- We could add the individual phones in the garbage model similar to the words
  - the weights could be trained on a development set
- Currently, the acoustic model used, has been trained on clean utterances
  - Can train it for the noisy campus data to make the overall system robust to noise
- Can add the dis-fluency path at phone level also to deal with sub-words
- Currently, we are manually building the pronunciation dictionary
  - the pronunciation model could be design to automate this process

List of Publications

1. K. Sabu, K. Kumar, and P. Rao " Improving the Noise Robustness of Prominence Detection for Children's Oral Reading Assessment", Proc. of NCC, Feb 2018, Hyderabad, India.

2. K. Sabu, K. Kumar, and P. Rao " Automatic detection of expressiveness in oral reading ", Show & Tell demonstration, Interspeech, Hyderabad, India, 2018.

3. P. Rao, M. Pandya, K. Sabu, K. Kumar, and N. Bondale " A Study of Lexical and Prosodic Cues to Segmentation in a Hindi-English Code-switched Discourse ", Interspeech, Hyderabad, India, 2018.

Thank You!

Questions?