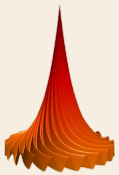
A study on native American English speech recognition by Indian listeners with varying word familiarity level

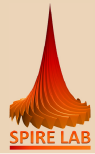
Abhayjeet Singh[1], Achuth Rao MV[1], Rakesh Vaideeswaran[1], Chiranjeevi Yarra[2], Prasanta Kumar Ghosh[1]

[1]SPIRE LAB, Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India
[2]Language Technologies Research Center (LTRC),
IIIT Hyderabad, India

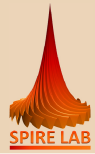


COCOSDA 2021

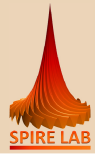
Overview



Overview

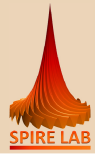


Non-Native listener and English as L2 Speech



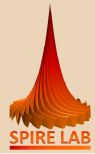
- Its challenging for a non-native English listener to recognize speech from a native English speaker.
- Recognition performance depends on the listener's experience or exposure to English language.

Motivation



- Ability to recognize English speech is crucial for understanding various online contents such as Massive Open Online Courses (MOOCs).
- Quantifying the difficulty of recognizing speech from native American English speakers is relevant in the Indian context.

Previous Work

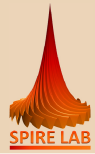


- Studies show that native listeners are better at recognizing the native speech compared to non-native speech in quiet and low noise conditions.[1]
- Some studies compare listeners in different noise types such as white, babble, pink, speech-shaped noise, etc.
- Effect of semantic context on target word recognition has also been studied.[2]
- Some of the mentioned studies also analyze the effect of listener's proficiency with a wide variety of target languages such as English, Dutch, Spanish, Swedish, Mandarin, Korean etc

[1] M. G. Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," The Journal of the Acoustical Society of America, vol. 119, no. 4, pp. 2445–2454, 2006.

[2] J. Aydelott, D. Baer-Henney, M. Trzaskowski, R. Leech,and F. Dick, "Sentence comprehension in competing speech: Dichotic sentence-word priming reveals hemispheric differences in auditory semantic processing," Language and Cognitive Processes, vol. 27, no. 7-8, pp.1108–1144, 2012.

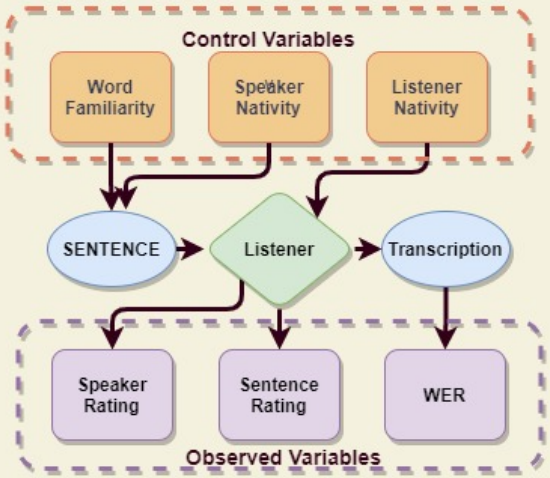
Objectives of the work



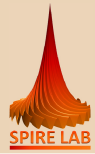
- We focus on the native American English speakers and listeners from Indian nativity.
- Hypothesize that the difficulty of recognizing words in a sentence depends on their frequency of occurrence (FoO-score).
- WF-score - sum of FoO score of the words in the sentence.
- Lower the WF-score of a sentence easier it is recognize and based on WF-score, sentences are categorized into three Word Familiarity Levels (WFLs): easy, medium and hard.

Proposed Strategy

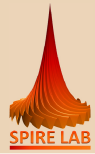
- This study mainly contains three control variables: word familiarity, speaker's nativity and listener's nativity.
- Indian listeners provide two ratings for a sentence : difficulty to understand the sentence and speaker's accent. Listeners also transcribes the sentence.
- With these observed variables (2 ratings and a transcription) together with three control variables, we study the effect of different sets of control variables on these observed variables.



Overview



Materials Used in Study

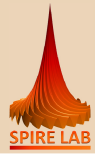


- In this study, 500 TIMIT[1] sentences were selected and the were listened to and recognized by a 500 Indian listeners of varied nativities. Each listener responded to subset of 10 sentences.
- TIMIT comprises 2342 unique sentences. To select 500 sentences, we use American National Corpus (ANC)[2] frequency dataset which consists $\sim$ 250K unique words ordered by the usage.

[1] L. D. Consortium et al., "The darpa timit acousticphonetic continuous speech corpus," NIST Speech CD, pp. 1–1, 1990.

[2] "American national corpus, second release, frequency data, last accessed:29/03/2021." [Online]. Available: https://www.anc.org/data/anc-secondrelease/frequency-data/

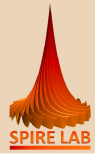
Data Collection



- 50 google forms were created using 500 categorized TIMIT sentences, each form containing 10 sentences : 5 easy, 3 medium and 2 hard category.
- Vocabulary of the selected 500 sentences: 1867 unique words (Easy: 910, Medium: 556, Hard: 401)
- Speaker's dialect region (DR) distribution for selected 500 TIMIT sentences over all three levels (Easy, Medium and Hard):

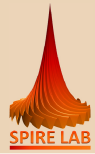
Dialect Region	DR1	DR2	DR3	DR4	DR5	DR6	DR7	DR8	Total
Easy	68	95	28	17	14	9	14	5	250
Medium	31	35	18	25	19	4	10	8	150
Hard	18	26	9	19	10	8	8	2	100
Total	117	156	55	61	43	21	32	15	500
TIMIT	490	1020	1020	1000	980	460	1000	330	6300

Observed Variables



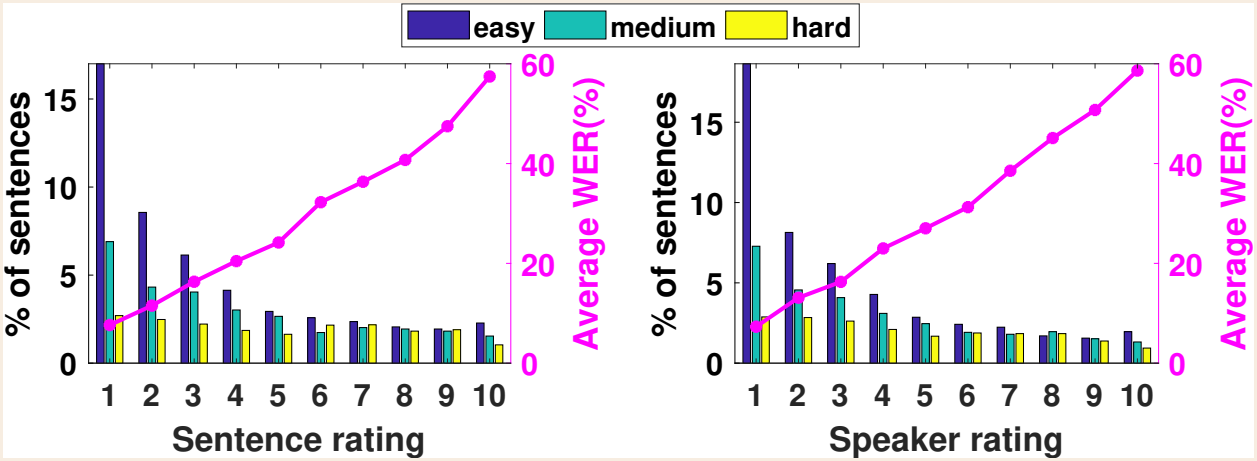
- In each form, respondent listens to the selected TIMIT sentence audio clip and then transcribes the sentence to the best of their ability.
- WER between the provided transcription and the original TIMIT sentence is calculated.
- Listeners then rate it on the basis of difficulty in recognizing the sentence and understanding the speaker.

Overview



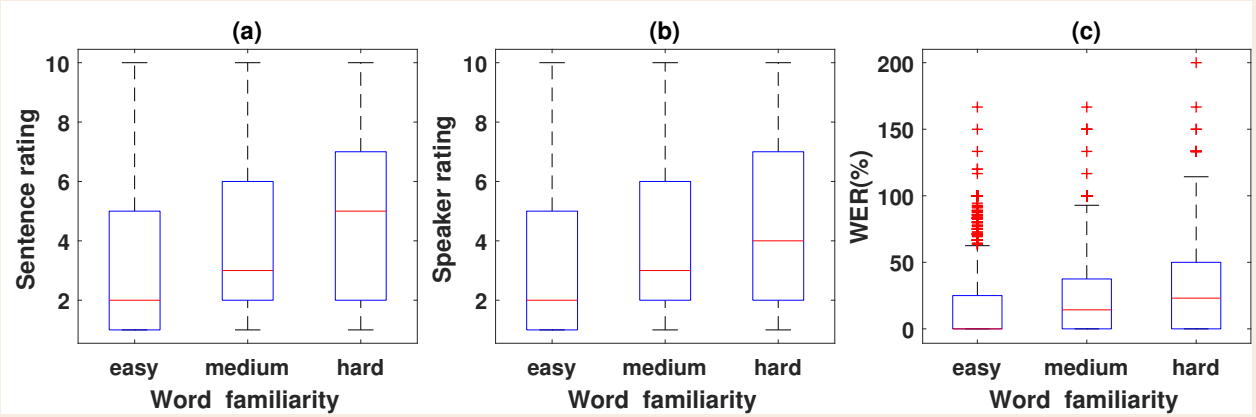
Relationship between Observed Variables

- Correlation of 0.90 between sentence difficulty and speaker accent difficulty ratings.
- WER increases as both the ratings go higher.
- Both the ratings provided by the listeners closely follow the three WFLs (Easy, Medium and Hard)



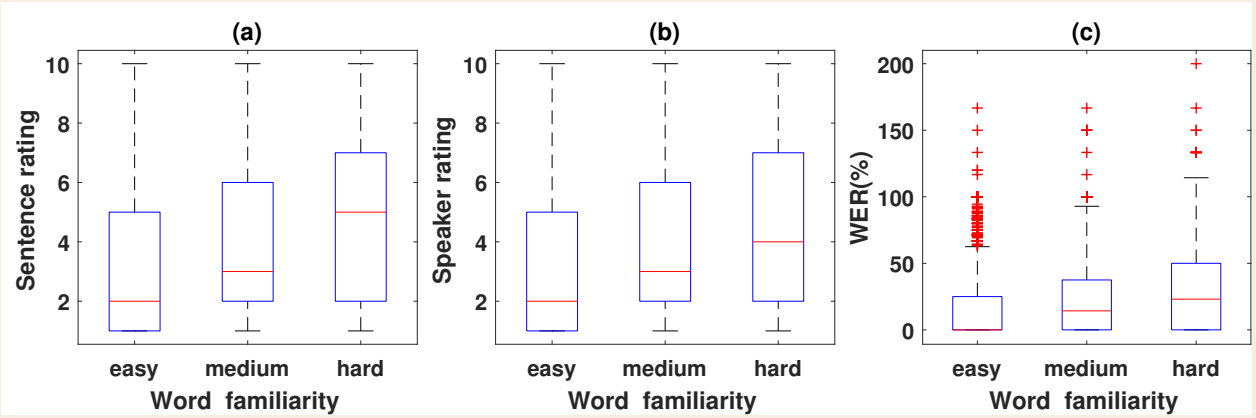
Effect of Word Familiarity

- Sentence difficulty ratings as well as the speaker difficulty ratings increase as WFL goes from easy to hard.
- Significant $(p < 0.01)$ increase of ratings from high to low WFL suggests that word familiarity plays a significant role in the perceived difficulty.



Effect of Word Familiarity

- WER(%) value increases with decreasing WFL.
- Percentages of zero WER for easy, medium and hard WFLs are 50.76%, 36.8% and 25.6% respectively, decline in percentages suggest that word familiarity significantly alters human speech recognition (HSR) performance.



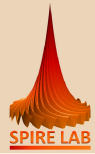
ASR vs HSR for different WFL

- Three ASRs were trained:
  - ASR1: Acoustic model (AM) for this ASR was trained on iTIMIT corpus[1]. Language model (LM) was trained on the TIMIT data.
  - ASR2: Both AM and LM were trained using LIBRI speech corpus[2] ($\sim$960 hrs)
  - ASR3: Same AM as of ASR2 but the LM was trained using both TIMIT and LIBRI speech text

[1] C. Yarra, R. Aggarwal, A. 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," in 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 2019, pp. 1–6.

[2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210.

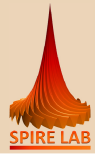
ASR vs HSR for different WFL



- WER for both HSR and ASR decreases with increase in WFL.
- ASR1 performs poorly for easy rather than medium/hard sentences.
- ASR2 performance is consistently higher than the HSR in all WFLs.

WER(%)	HSR	ASR 1 [iTIMIT]+{iTIMIT}	ASR 2 [LIBRI]+{LIBRI}	ASR 3 [LIBRI]+{LIBRI+iTIMIT}
Easy	17.43 (25.6)	29.70 (36.1)	13.21 (18.4)	2.82 (10.8)
Medium	23.17 (27.0)	21.05 (30.0)	20.67 (22.0)	6.37 (21.1)
Hard	30.78 (30.9)	28.47 (39.9)	26.57 (30.5)	10.02 (23.3)

ASR vs HSR for different WFL

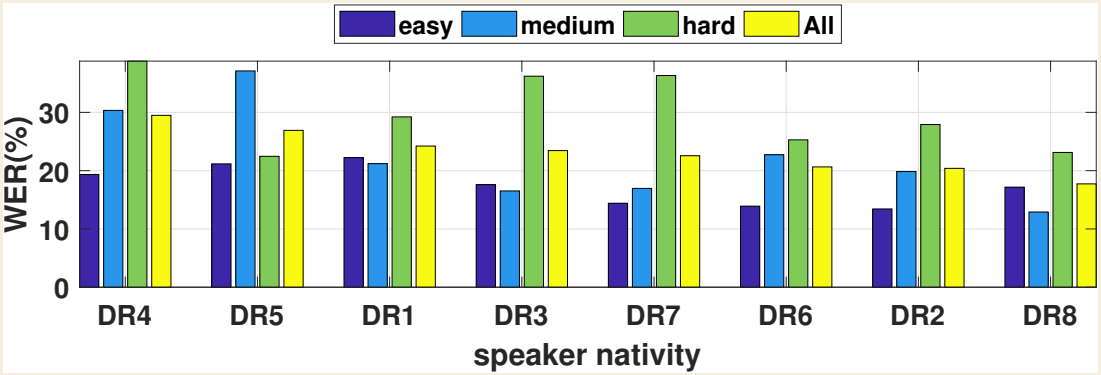


- Significant ($p < 0.01$) decrease in WER is observed as we shift from AM trained on Indian accent data to native American accent.
- ASR3, where LM is trained on both LIBRI and TIMIT, performance of ASR improves significantly $\rightarrow$ Inclusion of TIMIT data to LM.

WER(%)	HSR	ASR 1 [iTIMIT]+{iTIMIT}	ASR 2 [LIBRI]+{LIBRI}	ASR 3 [LIBRI]+{LIBRI+iTIMIT}
Easy	17.43 (25.6)	29.70 (36.1)	13.21 (18.4)	2.82 (10.8)
Medium	23.17 (27.0)	21.05 (30.0)	20.67 (22.0)	6.37 (21.1)
Hard	30.78 (30.9)	28.47 (39.9)	26.57 (30.5)	10.02 (23.3)

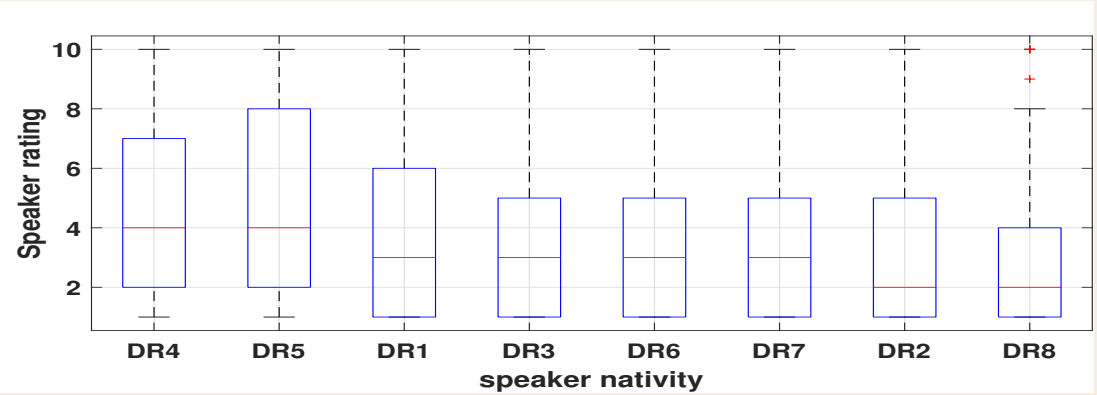
Impact of Speaker Nativity on Speech Recognition

- Higher the WFL lower the average WERs (except: DR8 & DR5).
- DRs at the extremities on the given plot are significantly different ($p < 0.01$)→Dialect region of the native American speakers is a key factor to influence the recognition accuracy by Indian listeners.

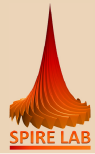


Impact of Speaker Nativity on Speech Recognition

- DR4 and DR5 have the highest speaker ratings and are significantly different $(p < 0.01)$ from rest of the nativities→Indian listeners' recognition accuracy varies with speaker's dialect.
- Indian listeners find speakers from DR8, DR2 and DR7 easier to follow significantly compared to DR1, DR4 and DR5.



Word Deviation and Familiarity



- Deviation of number of words between the transcript provided by the listener and the original sentence.
- Negative deviations are less than their positive counterparts.
- exact matches (word deviation=0) the percentage of deviations decrease with lowering WFL categories.
- DR4 and DR5 have the least percentages of sentences with exact matches→Indian listeners found American speakers from DR4 and DR5 as the most difficult ones to understand.

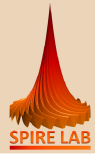
Deviation	**0**	**1**	**-1**	**2**	**-2**	> 2	< -2
Easy	75.16	10.8	4.72	2.88	0.8	5.48	0.16
Medium	67.87	10.6	8.93	3.87	1.07	7.27	0.4
Hard	58.3	13.6	13.2	3.4	2.1	8.8	0.6

WER vs Listener's Nativity

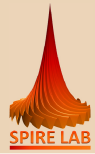
- Telugu which has the highest average WER is significantly different $(p < 0.01)$ from Kannada, Hindi and Tamil whereas Bengali which has second highest WER, is significantly different $(p < 0.01)$ from Hindi and Tamil.
- Could be due to the amount of exposure listeners of these nativities had to American English, in addition to the nativity specific factors in recognition.

	Telugu	Bengali	Kannada	Hindi	Tamil
WER	25.78 (26.94)	24.22 (40.36)	18.63 (23.76)	16.25 (24.43)	15.12 (20.65)

Overview



Conclusion



- Observed variables significantly increase with decrease in WFL.
- Observed variables vs Speaker's dialect→Speakers from DR4 and DR5 were found to be difficult to be followed by Indian listeners.
- Listeners' nativity plays a significant role in speech recognition.
- Performance of HSR and ASR1 were found to be similar. And ASR2 shows significant improvement over ASR1 is observed.
- Deviations in number of words rise with lowering WFL.

Acknowledgement

- Authors thank all the subjects for their participation in the study.

THANK YOU

Have Questions/Suggestions?
Write to us at spirelab.ee@iisc.ac.in