Towards Improving Intelligibility of Black-Box Speech Synthesizers in Noise

Thomas Manzini, Alan Black

Language Technologies Institute
School of Computer Science, Carnegie Mellon University

Introduction

- Research Question
  - Can we improve intelligibility of speech in noise without changing the synthesizer’s audio?

Introduction

- Research Question
  - Can we improve intelligibility of speech in noise without changing the synthesizer's audio?

- Motivation from the real world
  - Emergency personnel all need to be understood via radio
    - When asked to repeat a statement, speakers often rephrase their utterance to increase the chance of being understood
    - The speaker does this without knowing what he/she sounds like

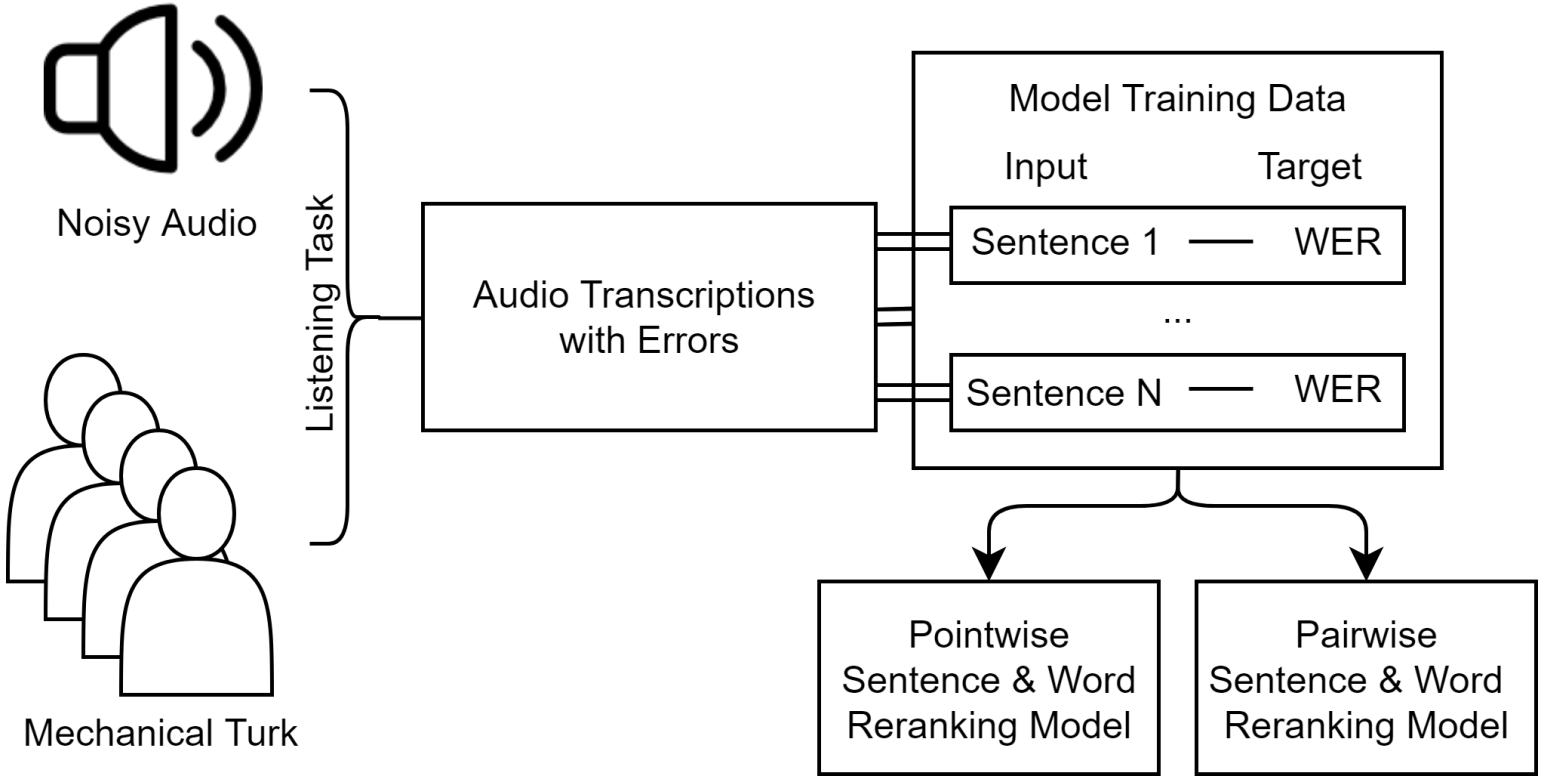
Related Work

- Work on intelligibility of speech in noise
  - The ability of a human to understand speech in noisy environments (a restaurant, on a battlefield, in a helicopter) has been studied in numerous ways.
  - The majority of these works approach this problem from a linguistic, psychological, or medical perspective.
  - Listening tests are the go-to methodology for determining what a human can and cannot understand in noise.

Related Work

- Work on estimating intelligibility of speech in noise
  - There has been some work on predicting the probability that a human will understand some given speech.
  - This work is limited and relies on measures of audio such as the glimpse proportion, the DAU metric, and others to act as features in statistical models.
  - To the best knowledge of the authors, no work has been done on estimating the intelligibility of speech in noise from non-audio based features.

Approach


Noisy Audio
Mechanical Turk
Listening Task
Audio Transcriptions with Errors
Model Training Data
Input
Target
Sentence 1
WER
...
Sentence N
WER
Pointwise Sentence & Word Reranking Model
Pairwise Sentence & Word Reranking Model

Listening Test Evaluation

- We evaluate user errors using Word Error Rate (WER), while other evaluations such as Concept Error Rate (CER) could have been more appropriate, it has been shown that WER closely follows CER.

- Based on a sample evaluation this appears to be the case for our data.

Approach

- We take a two step approach to this task.
  - First we collect data from a listening test performed on Amazon Mechanical Turk.
    - Users are asked to listen to 30 audio files and type the words that they hear spoken.
    - Users are presented with audio from 3 different synthesizers and played in 3 different noise settings.
    - One collection task each for training and testing data.

      Training: 45 Turkers, 450 audio files, 3 listeners per file
      Testing: 50 Turkers, 150 audio files, 10 listeners per file

Approach

- We take a two step approach to this task.
  - Second, we evaluate user performance and attempt to engineer features to predict intelligibility.
    - We explore predicting the intelligibility of individual words and of sentences as a whole.
    - We treat this problem as one of reranking and explore pointwise and pairwise reranking approaches.

Data

3 Synthesizers

E-Speak

Flite

Google

"Is the tv in the bathroom working properly?"

3 Noise Settings

Setting #1

"Is the phone in the living room ringing?"

Setting #2

"How warm is it in the dining room?"

Setting #3

"Are the den lights still on?"

Specifics of each noise setting and synthesizer are included in the paper

Listening Test Results

Noise levels are fairly different from each other.

Transcription Precision Score	Noise Level 1	Noise Level 2	Noise Level 3	Average
Espeak	0.227	0.196	0.242	0.222
Flite	0.346	0.375	0.343	0.355
Google	0.542	0.639	0.559	0.580
Average	0.372	0.403	0.381	

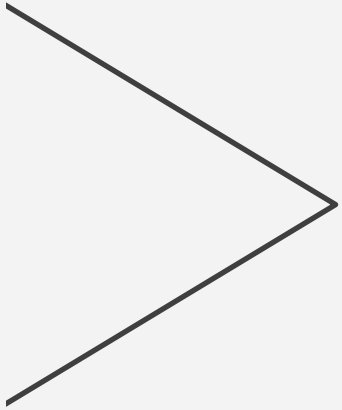
As the quality of your synthesizer increases, the more intelligible in noise it becomes.

Sentence Level Features & Model

- Average word rank
- Average word length
- Sentence length
- Word count
- Percent of unique characters

Sentence Level Features & Model

- Average word rank
- Average word length
- Sentence length
- Word count
- Percent of unique characters



Linear Model

Pointwise Sentence Reranking

MSE Looks great…
What's the problem?

Point Wise Reranking - Sentences			
Synthesizer	Noise Level	MSE (Test)	Spearman's R (Test)
Espeak	1	0.0227	0.2258
Espeak	2	0.0256	-0.3728
Espeak	3	0.0311	-0.4650
Flite	1	0.0477	-0.1225
Flite	2	0.0451	0.4621
Flite	3	0.0587	-0.1863
Google	1	0.0505	0.1176
Google	2	0.0915	-0.2943
Google	3	0.0701	0.0662

The model's predictions are do not correlate with the data overall

Predict the WER of from linguistic features of a sentence.

Pairwise Sentence Reranking

MSE is all over the place

Pair Wise Reranking - Sentence			
Synthesizer	Noise Level	MSE (Test)	MSE Variance (Test)
Espeak	1	0.3441	0.1211
Espeak	2	0.5960	0.1217
Espeak	3	0.4641	0.2591
Flite	1	0.7074	0.1496
Flite	2	0.3979	0.351
Flite	3	0.6095	0.3688
Google	1	0.6679	0.1638
Google	2	0.4076	0.1440
Google	3	0.5080	0.1966

Variance is higher than we would like as well

Predict if the WER of sentence one or two will be higher from linguistic features of a sentence.

Word Level Features

Word Level Features

- Word rank
- Percent of vowels in the word
- Percent of consonants in the word
- Length of the word
- Percent of unique characters in the word

Context Features

- Above features for the previous and next word
- Number of words in the sentence
- Number of unique words in the sentence

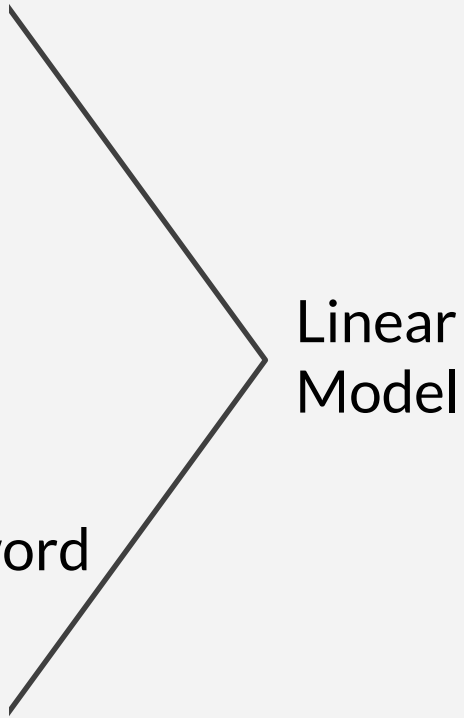
Word Level Features

Word Level Features

- Word rank
- Percent of vowels in the word
- Percent of consonants in the word
- Length of the word
- Percent of unique characters in the word

Context Features

- Above features for the previous and next word
- Number of words in the sentence
- Number of unique words in the sentence



Pointwise Word Reranking

MSE looks okay, but not fantastic…

Point Wise Reranking - Words			
Synthesizer	Noise Level	MSE (Test)	Spearman's R (Test)
Espeak	1	0.0429	-0.2516
Espeak	2	0.0426	0.0676
Espeak	3	0.0318	-0.1120
Flite	1	0.0932	-0.1612
Flite	2	0.1032	0.1346
Flite	3	0.0468	-0.1589
Google	1	0.0902	0.2173
Google	2	0.1182	0.3114
Google	3	0.0801	0.0178

But we are correlating better

Particularly for Google

Predict the WER of a specific word in a sentence using linguistic features.

Pairwise Word Reranking

MSE is much tighter than pairwise sentence level

Pair Wise Reranking - Words			
Synthesizer	Noise Level	MSE (Test)	MSE Variance (Test)
Espeak	1	0.1946	0.1151
Espeak	2	0.1546	0.0901
Espeak	3	0.1610	0.0959
Flite	1	0.1896	0.0972
Flite	2	0.2022	0.1167
Flite	3	0.2009	0.1052
Google	1	0.1783	0.0941
Google	2	0.1794	0.1031
Google	3	0.1840	0.1076

Variance is much lower than in the pairwise sentence level

Predict if the WER of word one or two will be higher from linguistic features of a sentence.

Discussion

- The pairwise setting worked the best overall, with the pairwise word level model working the best.

Discussion

- The pairwise setting worked the best overall, with the pairwise word level model working the best.

- We hypothesize that these results would stabilize with more data but additional experimentation is required.

Discussion

- The pairwise setting worked the best overall, with the pairwise word level model working the best.
- We hypothesize that these results would stabilize with more data but additional experimentation is required.
- From an error analysis perspective the largest source of error came from the results of the listening tests.

Future Work

- The most critical element needed is more data.

Future Work

- The most critical element needed is more data.
- Additional exploration of the different types of noise settings
  - We only explored a limited space of noise, and noise settings 1 and 3 ended up being similarly intelligible.
  - More appropriate noise settings for emergency response would also be relevant (rescue devices, engine noises, etc).

Future Work

- The most critical element needed is more data.
- Additional exploration of the different types of noise settings.
  - We only explored a limited space of noise, and noise settings 1 and 3 ended up being similarly intelligible.
  - More appropriate noise settings for emergency response would also be relevant (rescue devices, engine noises, etc).
- More experimentation with different synthesizers as well
  - We chose three synthesizers with a range of quality to demonstrate the effectiveness of this approach, but this approach may not work for all synthesizers.

Takeaways

- We are able to predict rank the intelligibility of specific words using a pairwise reranking setting.
  - Currently unable to rank intelligibility of words and sentences in the pointwise setting and sentences in the pairwise reranking setting.

Takeaways

- We are able to predict rank the intelligibility of specific words using a pairwise reranking setting.
  - Currently unable to rank intelligibility of words and sentences in the pointwise setting and sentences in the pairwise reranking setting.

- We believe that this approach shows promise and with additional labeled data from listening tests these ranking models could improve.

Acknowledgements

- Carolyn Rose
- Rajat Kulshreshtha
- Abhilasha Ravichander
- Officers of CMU EMS
- Elise Romberger

- All Paper Reviewers

Thank you!