Evaluation of Off-the-shelf Speech Recognizers on Different Accents in a Dialogue Domain

Divya Tadimeti[1,2], Kallirroi Georgila[2], David Traum[2]
[1]University of California, Berkeley
[2]Institute for Creative Technologies, University of Southern California

Language Resources and Evaluation Conference (LREC) 2022, Marseille, France

USC Institute for Creative Technologies

Motivation

- Recent evaluation of Automatic Speech Recognition (ASR) systems on speech directed at computer agents has shown that ASR systems are continuously getting better (Georgila et al., 2020)
- Recent work has shown that ASR systems have a much higher error rate on speakers of African American Vernacular English than on rural White Californians engaging in sociolinguistic interviews (Koenecke et al., 2020)
- It is an open research question whether the pattern observed by Koenecke et al. also holds for other kinds of accents and agent-directed speech

Outline

- Data
- Speech recognizers
- Results
- Conclusion

Data

- 2281 utterances collected between human participants and SGT Blackwell
- SGT Blackwell is a question-answering character developed at ICT who answers general questions about the Army, himself, and his technology
- Speech collected from visitors to the Cooper-Hewitt Museum in New York from December 2006 to March 2007

Data – example dialogue

Excerpt of an interaction between SGT Blackwell and a museum visitor:

Museum visitor: What is your favorite color?
SGT Blackwell: I like red, white, and blue.
Museum visitor: Why do you like red?
SGT Blackwell: I am not authorized to comment on that.

Data annotation with accent information

- We listened to every file to identify the accent of the speaker
- To measure inter-annotator agreement, 3 annotators (2 American native speakers of English, 1 non-native but fluent speaker of English) listened to 157 audio files
- 8 accent categories:
  - General American, Northeast American, British, Indian, French, East Asian, European uncategorized, non-American uncategorized

Examples – can you guess the accent?

General American, General American, American

British, non-American uncat, non-American uncat

Indian, Indian, Indian

European uncat, European uncat, European uncat

British, British, British

French, French, French

Northeast American, Northeast American, American

East Asian, East Asian, East Asian

Inter-annotator agreement

Annotators and labelling setup	Krippendorff's alpha	Absolute agreement (%)
Annotators 1, 2, 3 (American, British, Indian, French, East Asian, European Uncat & non-American Uncat)	0.719	76.43
Annotators 1, 2, 3 (American & Else)	0.879	95.33
Annotators 1, 2 (General American, Northeast American, British, Indian, French, East Asian, European Uncat & non-American Uncat)	0.672	71.34
Annotators 1, 2 (General American, Northeast American & Else)	0.8	91.72
Annotators 1, 2 (American, British, Indian, French, East Asian, European Uncat & non-American Uncat)	0.712	75.80
Annotators 1, 2 (American & Else)	0.9	96.18
Annotators 1, 3 (American, British, Indian, French, East Asian, European Uncat & non-American Uncat)	0.719	76.43
Annotators 1, 3 (American & Else)	0.835	93.63
Annotators 2, 3 (American, British, Indian, French, East Asian, European Uncat & non-American Uncat)	0.725	77.07
Annotators 2, 3 (American & Else)	0.901	96.18

Speech recognizers

ASR	Location	Type of processing	Model used
Amazon cloud online	cloud	online	
Apple device online	device	online	
Apple cloud online	cloud	online	
Google cloud online command_and_search	cloud	online	command_and_search
Google cloud online default	cloud	online	default
Google cloud online phone_call	cloud	online	phone_call
Google cloud online video	cloud	online	video
IBM cloud online	cloud	online	
Kaldi device offline ASpIRE	device	offline	ASpIRE
Kaldi device online ASpIRE	device	online	ASpIRE
Kaldi device offline LibriSpeech	device	offline	LibriSpeech
Kaldi device online LibriSpeech	device	online	LibriSpeech
Microsoft cloud offline	cloud	offline	
Microsoft cloud online	cloud	online	

Evaluation metric – word error rate

Word Error Rate = ((#Insertions + #Substitutions + #Deletions) / #Words_in_reference_transcription) x 100%

Example

Reference transcription: where were you born pal
ASR output: uh where are you born
#Insertions = 1
#Substitutions = 1
#Deletions = 1
#Words_in_reference_transcription = 5
Word Error Rate = 60%

Results

ASR	General American N=1767	Regional American N=96	All American N=1863	All Non-American N=418
Amazon cloud online	18	20.3	18.14	25.54
Apple device online	12.76	24.1	13.45	18.95
Apple cloud online	**10.21**	22.2	**10.95**	**12.52**
Google cloud online command_and_search	11.94	15.37	12.15	14.66
Google cloud online default	13.19	18.22	13.49	16.97
Google cloud online phone_call	14.06	15.18	14.13	16.31
Google cloud online video	11.24	**11.39**	11.25	14.61
IBM cloud online	26.93	28.65	27.04	33
Kaldi device offline ASpIRE	25.57	28.27	25.74	34.27
Kaldi device online ASpIRE	32.48	28.46	32.24	41.13
Kaldi device offline LibriSpeech	41.59	46.49	41.89	52.83
Kaldi device online LibriSpeech	45.15	46.87	45.26	56.67
Microsoft cloud offline	15.47	18.6	15.66	18.51
Microsoft cloud online	15.57	17.84	15.71	19.71

Results (continued)

ASR	Non-American Uncat N=162	European Uncat N=92	French N=39	British N=88	East Asian N=21	Indian N=16
Amazon cloud online	25.25	16.84	33.13	29.08	34.34	27.4
Apple device online	21.56	13.52	15.63	18.37	20.2	31.51
Apple cloud online	**12.77**	**6.89**	**5.63**	18.37	19.19	15.07
Google cloud online command_and_search	14.18	12.5	13.13	**17.09**	21.21	**12.33**
Google cloud online default	15.46	13.78	20.63	21.68	19.19	**12.33**
Google cloud online phone_call	13.05	11.73	23.75	22.7	23.23	**12.33**
Google cloud online video	12.91	13.52	11.88	18.11	21.21	15.07
IBM cloud online	34.75	22.96	46.88	35.71	31.31	27.4
Kaldi device offline ASpIRE	32.62	27.3	43.13	38.78	42.42	32.88
Kaldi device online ASpIRE	38.87	38.01	52.5	45.41	37.37	36.99
Kaldi device offline LibriSpeech	61.28	40.56	62.5	44.64	54.55	57.53
Kaldi device online LibriSpeech	66.81	45.92	63.75	46.94	57.58	52.05
Microsoft cloud offline	19.43	14.54	9.38	23.21	**15.15**	30.14
Microsoft cloud online	23.26	11.73	9.38	24.74	16.16	28.77

Conclusion

- The performance of the ASR systems for non-American accents is considerably worse than for General American accents
- Depending on the recognizer, the absolute difference in performance between General American accents and all non-American accents combined can vary approximately from 2% to 12%, with relative differences varying approximately between 16% and 49%
- This drop in performance becomes even larger when we consider specific categories of non-American accents
- There are performance differences across ASR systems, and while the same general pattern holds, with more errors for non-American accents, there are some accents for which the best recognizer is different than in the overall case

Thank you!

Questions?

Funding sources: National Science Foundation, Army Research Laboratory