Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6001–6008
Marseille, 20-25 June 2022

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]Electrical Engineering & Computer Science, University of California, Berkeley
[2]Institute for Creative Technologies, University of Southern California
dtadimeti@berkeley.edu, {kgeorgila,traum}@ict.usc.edu

Abstract

We evaluate several publicly available off-the-shelf (commercial and research) automatic speech recognition (ASR) systems on dialogue agent-directed English speech from speakers with General American vs. non-American accents. Our results show that 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 indicating a need for more diligent collection of and training on non-native English speaker data in order to narrow this performance gap. 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. We expect these results to be useful for dialogue system designers in developing more robust inclusive dialogue systems, and for ASR providers in taking into account performance requirements for different accents.

Keywords: speech recognition evaluation, accents, dialogue systems

1. Introduction

Automatic speech recognition (ASR) systems are being used for an increasing number of speech-to-text applications. With this proliferation, it is increasingly important for this technology to serve all subgroups of consumers. Recent work has shown that ASR systems have a much higher error rate on speakers of African American Vernacular English (AAVE) than on rural White Californians engaging in sociolinguistic interviews (Koenecke et al., 2020). Recent evaluation of ASR systems on speech directed at computer agents (Georgila et al., 2020) shows that speech recognizers have been getting better recently on agent-directed speech compared to previous years (Yao et al., 2010; Morbini et al., 2013), but leaves open the question of whether this performance is equivalent for different speakers or whether the pattern observed by (Koenecke et al., 2020) also holds for other kinds of accents and for agent-directed speech.

To this end, we evaluate the performance of popular ASR platforms — Google, Microsoft, Apple, Amazon, IBM, and Kaldi — on English speech from populations with different accents. We begin with a high-level distinction between General American accents and non-American accents, and then focus on more specific categories of non-American accents including French, Indian, British, and East Asian accents. We report on a re-analysis of a subset of the ASR outputs examined by Georgila et al. (2020), including ASR outputs using 2 additional configurations of the Google ASR platform that were not reported in (Georgila et al., 2020), and new annotations for speaker accent. This is also an extension of our recent work (Tadimeti et al., 2021), where we reported on preliminary results on ASR performance for different accents using a subset of the ASR systems and configurations that we use here.

2. Data

We evaluated the ASR systems on a dataset of 2281 utterances collected from conversation between human participants and SGT Blackwell, a virtual agent developed by the USC Institute for Creative Technologies. SGT Blackwell (Leuski et al., 2006) is a question-answering character who answers general questions about the Army, himself, and his technology. Speech comes from visitors to the Cooper-Hewitt Museum in New York from December 2006 to March 2007, who interacted with SGT Blackwell at his booth as part of the National Design Triennial exhibition (Robinson et al., 2008).

SGT Blackwell is designed to answer independent questions. The questions collected from sessions with SGT Blackwell come from the general public. The museum exhibit listed a set of about five sample questions, but visitors were free to ask anything they wanted. The following utterances illustrate a segment of a dialogue between a museum visitor and SGT Blackwell:

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.

In this museum setting open to the general public, it was assumed that the majority of visitors would be

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

Table 1: Krippendorff's alpha values and absolute agreement percentages for different comparisons. "American" means that the General American and Northeast American accents are merged into one category. "Else" means that all non-American accents are merged into one category. Note that Annotator 3 did not distinguish between General American and Northeast American accents, and annotated those instances as one "American" category.

American native speakers of English. Thus, the ASR component of SGT Blackwell used acoustic models for American English. Similar to this setup, in our experiments below we use commercial and research ASR systems with default settings for American English accents.

Speakers were anonymous and not identified in the data. In order to categorize the speech by accent, we listened to every audio file. Using this method, we manually classified the audio files into two main groups: General American English and non-American English accents. We use the term "General American" to encompass the utterances in our dataset lacking distinct regional and social characteristics (Wells, 1982; Van Riper, 1986). This includes mostly Western and Midwestern English accents and excludes noticeably Northeastern accents (i.e., New York, Boston), Southern American accents, and distinct dialects such as AAVE. Next, we segmented the non-American subset further into subcategories of non-American accents, the most common of which in our dataset were French, British, Indian, and East Asian. In some cases, it was not possible to distinguish the precise accent, so we also included an "uncategorized" class. For each non-American subset of files, we grouped utterances by individual speakers for additional analysis. These categories were then used to compute category-specific error rates for the recognizer results reported in (Georgila et al., 2020).

To assess inter-annotator reliability of accent classification, three annotators listened to a subset of 157 audio files and annotated the accent in each file as General American, Northeast American, British, Indian, French, East Asian, European uncategorized, and non-American uncategorized (8 distinct categories). Two of the annotators (Annotators 1 and 2) were American native speakers of English, and the third annotator was a native speaker of Greek but fluent speaker of English (Annotator 3). Note that Annotator 3 did not distinguish between General American and Northeast American accents, and annotated those instances as one "American" category. Agreement results between annotators are shown in Table 1. Krippendorff's alpha between Annotators 1 and 2 was measured at 0.672 (with absolute agreement at 71.34%) when all 8 distinct categories were considered. Krippendorff's alpha among all 3 annotators was measured at 0.719 (with absolute agreement at 76.43%) when General American and Northeast American accents were merged into one "American" category. We also calculated pairwise inter-rater agreement scores after merging the General American and Northeast American accents into one "American" category, and after merging all non-American accents into one large category "Else". The results shown in Section 4 are based on the annotations of Annotator 1.

3. Speech Recognizers

The following publicly available ASR platforms were used in our evaluation: Amazon, Apple, Google, IBM, Kaldi, and Microsoft. All are commercial platforms except for Kaldi which has been developed in academia. We used a subset of the ASR outputs examined by Georgila et al. (2020), including ASR outputs using 2 additional configurations of the Google ASR platform that were not reported in (Georgila et al., 2020). Below we provide more details about the setup of each of these platforms used in our experiments.

Most of our testing of commercial ASR systems was done in online mode on an iPhone; for more details see Georgila et al. (2020). We streamed audio to the ASR services in 0.1 second chunks at 0.1 intervals simulating a user talking into a microphone. We have also done some limited testing in offline mode where we submitted each audio file to the ASR services in one chunk. In general, we expect ASR in offline mode to perform better than in online as it has all of the audio available to it at the same time, but in practice this is not always the case. In contrast to the commercial speech recognition platforms we conducted our Kaldi experiments on a local desktop machine.

3.1. Amazon

Amazon provides ASR under the name of Amazon Transcribe[1]. The iOS SDK (software development kit) is available on GitHub[2]. The SDK requires an AWS account with appropriate privileges for accessing the Transcription service. At the time of testing, the service was free for 60 minutes for the first 12 months and $0.024 per minute afterwards.

3.2. Apple

Apple provides ASR as a part of the Speech Framework included with both iOS and macOS. The ASR has both cloud and on-device options. At the time of testing, the cloud access was free, however Apple limited the number of requests to the cloud-based ASR from a single device per hour (1000), and the length of the audio for each request (< 1 min). The on-device recognition option had no limitations. In this study we used both the cloud-based ASR and the on-device ASR running on iPhone XS.

3.3. Google

Google provides ASR as a part of the Google Cloud platform under the name Cloud Speech-to-Text[3]. The SDK is available on GitHub[4] and requires a Google Cloud account. Google offers several pre-built ASR models, i.e., for phone call transcription (phone_call), short queries (command_and_search), video transcription (video), and one model for the other types of speech (default). At the time of testing, the service was free for the first 60 minutes and $0.024 or $0.036 per minute afterwards depending on the ASR model used. In this study we used the video, default, phone_call, and command_and_search models. In (Georgila et al., 2020) we only used the video and default models.

[1] https://aws.amazon.com/transcribe/

[2] https://github.com/aws-amplify/aws-sdk-ios

[3] https://cloud.google.com/speech-to-text/

[4] https://github.com/GoogleCloudPlatform/ios-docs-samples

3.4. IBM

IBM ASR is a part of the Watson platform[5]. The iOS SDK is available in source form from GitHub[6]. To access the speech-to-text service, the API requires a token that can be obtained by setting up an IBM Cloud account and enabling the service via the web-based interface. At the time of testing, the first 500 minutes per month were free and between $0.02 and $0.01 per minute afterwards depending on the usage.

3.5. Kaldi

Kaldi is a state-of-the-art open-source ASR toolkit developed to support research in speech recognition (Povey et al., 2011). For our experiments we used the ASpIRE and LibriSpeech models.

The ASpIRE model is trained on the Fisher English corpus of conversational speech which has been augmented with impulse responses and noises to create multi-condition training. The Fisher English corpus consists of 16-bit 8kHz telephone speech so for our experiments we had to downsample our audio files from 16-bit 16kHz to 16-bit 8kHz.

The LibriSpeech model is trained on the LibriSpeech corpus, which is a large (1000 hour) corpus of English read speech derived from audio books in the LibriVox project. Speech is sampled at 16kHz, and the accents included in the corpus are various and not marked, with the majority being US-English.

Both models are available on the Kaldi website[7]. ASpIRE is a nnet3 chain model and LibriSpeech is a nnet2 chain model. A chain model is a type of DNN-HMM model. For LibriSpeech we used the pruned 3-gram language model. We also experimented with larger language models but this resulted in very slow processing because of extreme memory requirements.

3.6. Microsoft

Microsoft provides ASR as a part of the Azure platform under the name Cognitive Services: Speech-to-Text[8]. The SDK is available as a binary download from the company with code samples located on GitHub[9]. The SDK requires an Azure account. At the time of testing, the service was free for the first 300 minutes each month and $0.016 per minute afterwards. We ran the system in both offline and online modes.

3.7. Summary

Table 2 provides a summary of each one of the configurations that we used. Kaldi is only available to run on

[5] https://www.ibm.com/cloud/watson-speech-to-text

[6] https://github.com/watson-developer-cloud/swift-sdk

[7] https://kaldi-asr.org/models.html

[8] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

[9] https://github.com/Azure-Samples/cognitive-services-speech-sdk

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	

Table 2: ASR platforms and configurations used in our experiments.

a device. The rest of the ASR systems run on the cloud, except for the Apple one which also runs on a device.

4. Results

Our evaluation metric is word error rate (WER), a standard measure of ASR performance, used for example by both Koenecke et al. (2020) and Georgila et al. (2020). WER is calculated by comparing the ASR output to the reference manual transcription of what the speaker says. To measure the WER, we have to add the number of insertions (words that the ASR outputs but the speaker has not uttered), deletions (words that the speaker has uttered but the ASR does not output), and substitutions (words uttered by the speaker being replaced by other words in the ASR output), and then divide by the total number of words in the reference transcription. Thus WER is formulated as:

$$\text{WER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Length of reference string}} \times 100\%$$

We report on two types of WER: (1) The standard WER where we add the number of insertions, deletions, and substitutions for the whole corpus and then divide by the number of words in the reference transcriptions for the whole corpus. (2) The average WER per utterance where we calculate the WER per utterance and then divide by the number of utterances in our corpus. A large number of errors (insertions, deletions, substitutions) in an utterance may substantially increase the average WER per utterance, especially for a small number of utterances, but may have a lesser effect on the standard WER where errors are summed over the whole corpus. In Tables 3 and 4 we can see our results (standard WER and average WER per utterance respectively) for General American, regional American (e.g., Northeastern American, Southern American), all American (General American and regional American combined), and all non-American accents combined. Tables 5 and 6 show results (standard WER and average WER per utterance respectively) for non-American uncategorized, European uncategorized, French, British, East Asian, and Indian accents. Each table also shows the number of utterances in our dataset for each category.

The best ASR system varies depending on the accent. For General American accents Apple cloud online outperforms the rest, followed by Google cloud online video. For regional American accents, the best recognizer is Google cloud online video. Apple cloud online is also the best ASR system for all non-American accents combined, again followed by Google cloud online video. This is true for both the standard WER and the average WER per utterance.

For non-American accents the situation is similar. In terms of standard WER, Apple cloud online and Google cloud online video compete for the best performance depending on the accent, with Microsoft cloud online outperforming or performing similar to Google (but not Apple) for French and European uncategorized accents. Microsoft has the best performance for East Asian accents, and Google has the best performance for British and Indian accents. Overall, there are minor variations depending on whether we use the standard WER or the average WER per utterance. Of course for non-American accents we do not have many data points but it is clear that Google, Apple, and Microsoft have an advantage over the rest of the recognizers.

Not surprisingly the Kaldi LibriSpeech model produces high WERs given the fact that it has been trained on data from audio books, which are rather different from conversational speech. The ASpIRE model performs better than the LibriSpeech model due to the fact that it has been trained on conversational speech. However, it was trained on telephone speech which may have negatively affected its performance. Also, while the ASpIRE language model performs certainly better than the LibriSpeech language model for our purposes, it still

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

Table 3: Results in terms of standard WER (%). N shows the number of utterances considered per category.

ASR	General American N=1767	Regional American N=96	All American N=1863	All Non-American N=418
Amazon cloud online	19.16	18.29	19.12	24.82
Apple device online	13.88	24.59	14.44	17.14
Apple cloud online	**11.15**	22.43	**11.73**	**11.54**
Google cloud online command_and_search	15.03	15.81	15.07	14.45
Google cloud online default	16.39	17.62	16.45	16.94
Google cloud online phone_call	17.41	16.87	17.38	16.91
Google cloud online video	14.64	**13.34**	14.57	14.45
IBM cloud online	31.82	27.63	31.61	33.18
Kaldi device offline ASpIRE	30.78	27.93	30.63	34.28
Kaldi device online ASpIRE	39.54	28.78	38.99	42.7
Kaldi device offline LibriSpeech	50.49	46.51	50.29	56.79
Kaldi device online LibriSpeech	54.62	47.22	54.24	61.96
Microsoft cloud offline	16.37	17.48	16.42	16.43
Microsoft cloud online	16.27	19.87	16.45	17.13

Table 4: Results in terms of average WER (%) per utterance. N shows the number of utterances considered per category.

generates errors that could have been avoided with a more extensive language model. For example, in many cases it would output the correct word but not in the same form as in the reference transcription (e.g., 'fail' vs. 'failed'), and this would increase the WER. As expected, in most cases, the Kaldi and Microsoft offline models performed better than their online counterparts. This is because in offline mode the ASR has all of the audio available to it at the same time. In most cases, Google cloud online phone_call has worse performance than the other Google variants because it is designed for audio over the phone.

5. Discussion

Most ASR systems perform fairly well for General American accents, but all of them do considerably worse for non-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, e.g., French, British, East Asian, etc.

The performance gap suggests that consumers with non-American English accents may find it considerably harder to take advantage of speech recognition technology. It is an open research question and an active area of research whether speech recognizers should be expected to perform equally well for native and

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

Table 5: Results in terms of standard WER (%). N shows the number of utterances considered per category.

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	24.43	18.6	30.58	28.94	28.28	23.13
Apple device online	18.19	14.14	15	16.88	17.53	29.9
Apple cloud online	**11.52**	**5.77**	7.01	17.26	16.51	17.92
Google cloud online command_and_search	13.98	12.23	14.74	**16.37**	17.76	**16.35**
Google cloud online default	15.53	13.39	21.71	21.33	16.64	**16.35**
Google cloud online phone_call	13.33	12.73	24.4	23.52	21.53	**16.35**
Google cloud online video	12.49	14.14	10.6	17.83	20.04	19.48
IBM cloud online	34.14	22.66	47.46	37.1	29.93	31.84
Kaldi device offline ASpIRE	31.53	27.92	40.37	42.65	40.15	30.11
Kaldi device online ASpIRE	38.55	41.8	50.97	50.18	37.02	35.99
Kaldi device offline LibriSpeech	67.42	43.61	63.01	48.07	51.42	64.82
Kaldi device online LibriSpeech	73.67	52.81	64.63	50.05	50.74	69.77
Microsoft cloud offline	18.28	13.47	**6.45**	20.13	**15.65**	19.79
Microsoft cloud online	18.9	11.83	9.79	21.97	19.29	18.23

Table 6: Results in terms of average WER (%) per utterance. N shows the number of utterances considered per category.

non-native speakers of a language, in our case American English (Le et al., 2007; Ghorbani and Hansen, 2018; Jain et al., 2018; Viglino et al., 2019; Ahamad et al., 2020; Shibano et al., 2021; Sullivan et al., 2022; Tong et al., 2022). Nevertheless, the performance gap shown in our results is wide enough to suggest that there is potential for improvement. To improve performance, ASR systems should be trained on more diverse speaker data (Fukuda et al., 2018). This requires more diligent collection of non-American English speaker data.

The above analysis is still preliminary in several respects. We are currently analyzing the errors of individual speakers and also calculating the impact of these speech errors on agent response selection. We would like to enhance the analysis by looking at additional domains of agent-directed speech and additional demographic groups (such as regional American accents, gender, age, etc.). Additionally, we would like to attempt a more objective approach to accent categorization, e.g., using databases such as eWAVE to make more linguistically-informed data categorizations, or analyzing or collecting data with demographic information about the speakers.

6. Conclusion

We evaluated several publicly available off-the-shelf (commercial and research) ASR systems on dialogue agent-directed English speech from speakers with General American vs. non-American accents. We found that the performance of the ASR systems for non-American accents is considerably worse than for General American accents. Our results indicate a need for more diligent collection of and training on non-native English speaker data in order to narrow this performance gap. 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. We expect these results to be useful for dialogue system designers in developing more robust inclusive dialogue systems, and for ASR providers in taking into account performance requirements for different accents.

7. Acknowledgments

The first author was supported by the NSF REU program, award 1852583, during her internship at the USC Institute for Creative Technologies. The second and third authors were sponsored by the U.S. Army Research Laboratory (ARL). Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

8. Bibliographical References

Ahamad, A., Anand, A., and Bhargava, P. (2020). AccentDB: A database of non-native English accents to assist neural speech recognition. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 5351–5358, Marseille, France (online).

Fukuda, T., Fernandez, R., Rosenberg, A., Thomas, S., Ramabhadran, B., Sorin, A., and Kurata, G. (2018). Data augmentation improves recognition of foreign accented speech. In *Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2409–2413, Hyderabad, India.

Georgila, K., Leuski, A., Yanov, V., and Traum, D. (2020). Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, pages 6469–6476, Marseille, France (online).

Ghorbani, S. and Hansen, J. H. (2018). Leveraging native language information for improved accented speech recognition. In *Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2449–2453, Hyderabad, India.

Jain, A., Upreti, M., and Jyothi, P. (2018). Improved accented speech recognition using accent embeddings and multi-task learning. In *Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2454–2458, Hyderabad, India.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 117(14):7684–7689.

Le, J. T., Best, C. T., Tyler, M. D., and Kroos, C. (2007). Effects of non-native dialects on spoken word recognition. In *Proceedings of the 8th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 1589–1592, Antwerp, Belgium.

Leuski, A., Patel, R., Traum, D., and Kennedy, B. (2006). Building effective question answering characters. In *Proceedings of the 7th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 18–27, Sydney, Australia.

Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., and Traum, D. (2013). Which ASR should I choose for my dialogue system? In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pages 394–403, Metz, France.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (2011). The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA.

Robinson, S., Traum, D., Ittycheriah, M., and Henderer, J. (2008). What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 1125–1131.

Shibano, T., Zhang, X., Li, M. T., Cho, H., Sullivan, P., and Abdul-Mageed, M. (2021). Speech technology for everyone: Automatic speech recognition for non-native English with transfer learning. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP)*, online.

Sullivan, P., Shibano, T., and Abdul-Mageed, M. (2022). Improving automatic speech recognition for non-native English with transfer learning and language model decoding. In *arXiv*.

Tadimeti, D., Georgila, K., and Traum, D. (2021). How well can an agent understand different accents? In *5th Widening NLP (WiNLP) Workshop - Co-located with EMNLP*, Punta Cana, Dominican Republic.

Tong, F., Li, T., Liao, D., Xia, S., Li, S., Hong, Q.,

and Li, L. (2022). The XMUSPEECH system for accented English automatic speech recognition. *Applied Sciences*, 12(1478).

Van Riper, W. R. (1986). General American: An ambiguity. *Dialect and Language Variation*, pages 123–135.

Viglino, T., Motlicek, P., and Cernak, M. (2019). End-to-end accented speech recognition. In *Proceedings of the 20th Annual Conference of the Speech Communication Association (INTERSPEECH)*, pages 2140–2144, Graz, Austria.

Wells, J. C. (1982). *Accents of English, Volume 3: Beyond the British Isles*. Cambridge University Press.

Yao, X., Bhutada, P., Georgila, K., Sagae, K., Artstein, R., and Traum, D. (2010). Practical evaluation of speech recognizers for virtual human dialogue systems. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 1597–1602, Valletta, Malta.