Evaluation of Off-the-shelf Whisper Models for Speech Recognition Across Diverse Dialogue Domains

Kallirroi Georgila and David Traum

Abstract We evaluate various off-the-shelf multilingual and English-only OpenAI Whisper models for automatic speech recognition (ASR) across diverse dialogue domains in American English. We also compare the performance of the above models on speech from speakers with General American versus non-American accents. We discuss our results in relation to our previous work in which we used various off-the-shelf commercial and research ASR systems. Our evaluation is targeted at non-experts with limited experience in ASR, and we expect it to be useful for dialogue system designers and ASR consumers who are trying to decide whether to switch from a commercial speech recognizer (e.g., Google, Apple, Microsoft) to a Whisper model.

1 Introduction

Automatic speech recognition (ASR) is an important component of a spoken dialogue system. Because other natural language processing modules process the ASR output, ASR directly affects overall system performance. In this paper we evaluate multiple off-the-shelf OpenAI Whisper models for ASR, using data collected from deployed spoken dialogue systems as well as from human-human conversations in 5 domains (6 data sets) in American English. This is our fourth large-scale ASR evaluation using corpora from a variety of domains [35, 20, 11]. In our third and most recent large-scale evaluation [11], we used multiple off-the-shelf state-of-the-art publicly available speech recognizers, both commercial (Amazon, Apple, Google, Microsoft, IBM) and research (Kaldi [21]). Here we evaluate the same data

Kallirroi Georgila, USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 USA, e-mail: kgeorgila@ict.usc.edu
David Traum, USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 USA, e-mail: traum@ict.usc.edu

sets as in [11] but our focus is on Whisper ASR performance. We are not aware of other large-scale evaluations of Whisper for dialogue domains.

In addition to evaluating the performance of multiple Whisper models for ASR on a diverse set of dialogue domains, we also compare the performance of these models on speech from speakers with General American versus non-American accents. This continues our previous work on evaluating the performance of off-the-shelf commercial and research ASR systems on different accents. In [27, 28] we evaluated the performance of Google, Microsoft, Apple, Amazon, IBM, and Kaldi, on English speech from populations with different accents. Here we focus on Whisper ASR and we are particularly interested in comparing the performance of multilingual versus English-only models.

The remainder of the paper describes related work, the data used, Whisper and its models, and the results of our evaluation. We also discuss the results in relation to our previous work. Finally we conclude and present possible directions for future work. Similar to [11, 27, 28] our evaluation is targeted at non-experts with limited experience in ASR, and we expect it to be useful for dialogue system designers and ASR consumers who are trying to decide whether to switch from a commercial speech recognizer (e.g., Google, Apple, Microsoft) to a Whisper model. Given that the previous models were evaluated 3–4 years ago, a fairer comparison of Whisper with other commercial and research ASR systems would use the latest versions of these recognizers, but this is left for future work. Nevertheless we believe that the experiments presented below will provide important insights into the strengths and weaknesses of Whisper and its off-the-shelf models for ASR for dialogue systems.

2 Related Work

In one of the earliest studies on ASR evaluation, Devine et al. [8] compared 3 commercial ASR software packages on medical progress notes and discharge summaries dictated by physicians. These ASR systems were IBM ViaVoice 98 with General Medicine Vocabulary; Dragon Systems NaturallySpeaking Medical Suite, version 3.0; and L&H Voice Xpress for Medicine, General Medicine Edition, version 1.2. The IBM system performed the best.

In a recent study, also in a medical domain, Kim et al. [15] tested 5 ASR platforms (Google Cloud, IBM Watson, Microsoft Azure, Trint, YouTube) in terms of transcription quality. The data used for testing were collected from the interaction of medical students with simulated patients. Note that simulated patients are human actors trained to act as patients in a medical situation. As expected, manual transcriptions were significantly more accurate than automatic transcriptions. Also, among the ASR systems, the automatic transcriptions of YouTube Captions significantly outperformed the other ASR platforms.

Broughton [6] evaluated 2 commercial ASR systems on conversational speech. The focus of this work was not so much on comparing ASR systems but on measuring speech recognition performance on conversational speech.

Burger et al. [7] evaluated 3 commercial desktop dictation ASR engines in 8 languages (US-English, UK-English, Iberian Spanish, French, German, Japanese, Simplified Chinese, and Traditional Chinese). ASR performance was better on read speech than spontaneous speech. Also, the ASR systems for US-English, Japanese, and Spanish performed better than the ASR systems for UK-English, German, French, and Chinese.

Gaida et al. [10] compared open-source speech recognizers from the Cambridge HTK family (HDecode v3.4.1, Julius v4.3), the CMU Sphinx family (Sphinx 4, PocketSphinx v0.8), and Kaldi. The evaluation was performed on the Verbmobil corpus (conversational speech in German) and the Wall Street Journal corpus (read speech in English). Gaida et al. [10] trained their own acoustic and language models for each corpus. The focus of this evaluation was on the ratio of effort (in setting up the toolkit for a specific corpus) to performance. Kaldi performed the best, and also provided easy to use training and decoding pipelines, and the most advanced techniques out of the box. Sphinx and HTK had comparable performance. However, for HTK to reach the performance of Sphinx, extensive effort was required on fine-tuning.

Këpuska and Bohouta [14] compared 2 commercial speech recognizers (Microsoft Speech API and Google Speech API) with an open-source speech recognizer (Sphinx 4), using audio files from the TIMIT speech database and the ITU (International Telecommunication Union). Google Speech API performed the best.

Baumann et al. [5] measured the overall accuracy and incremental performance of 2 open-source speech recognizers (Sphinx 4 and Kaldi) and a commercial speech recognizer (Google). Google performed the best in terms of overall accuracy. However, Google also exhibited a tendency to filter out disfluencies, which can be important information for incremental speech processing.

Addlesee et al. [1] evaluated 3 commercial ASR engines (IBM, Google, Microsoft), in an incremental setting focusing on the robustness of these ASR systems on conversational speech characteristics such as disfluencies and overlaps. They also evaluated the speaker diarization capabilities of these speech recognizers. They found that Microsoft was more robust to preserving speech disfluencies, IBM was more robust to preserving speech overlaps, and Google struck a balance between the two. But overall none of these ASR systems was suitable for reliable real-time conversational speech recognition.

Very recently Whetten et al. [34] evaluated 6 speech recognizers (2 cloud-based and 4 local) in an incremental spoken dialogue system setting. The cloud ASR systems were Google Cloud and Microsoft Azure. The local ones were Wav2Vec2, DeepSpeech, PocketSphinx, and Vosk. Whetten et al. [34] also evaluated an incremental version of RASA for natural language understanding (NLU). Note that RASA is a framework for NLU and developing conversational systems.

Over the years we have performed 3 large-scale ASR evaluations using much more diverse data sets and domains than previous work in the literature [35, 20, 11]. We have also employed a larger variety of ASR systems (both commercial and research) compared to previous work.

Our first large-scale ASR evaluation was done in 2010 [35]. We compared open-source speech recognizers from 2 main families: the Cambridge HTK family (HVite v3.4.1, HDecode v3.4.1, Julius v4.1.2) and the CMU Sphinx family (Sphinx 4, PocketSphinx v0.5). We tested these 5 ASR systems on data from 6 different dialogue domains. In this study, we did not focus on out-of-the-box models but instead trained our own acoustic and language models. Our results showed large differences in the recognition rates for the different domains, and for some domains the error rates were very high. Also, none of the 5 ASR systems dominated on all data sets.

Our second large-scale ASR evaluation was done in 2013 [20]. This evaluation included 2 research platforms, i.e., PocketSphinx, Otosense-Kaldi (a system developed at USC based on the research toolkit Kaldi [21]), and 3 commercial platforms, i.e., Apple Dictation, Google Speech API, AT&T Watson. This evaluation was an extension of our 2010 evaluation and included commercial cloud-based ASR services that achieved very good performance showing an absolute improvement of approximately 12%. But similarly to our first evaluation, none of the speech recognizers dominated on all data sets and there was large variation in performance depending on the domain.

Our third and most recent large-scale evaluation was done in 2020 and was aimed at non-experts with limited experience in ASR [11]. For this reason, we used various state-of-the-art publicly available speech recognizers, both commercial, i.e., Amazon, Apple, Google, Microsoft, and IBM, and research, i.e., Kaldi [21]. Our results showed major progress in ASR technology in the last few years especially with the use of deep learning techniques, and that the performance of each ASR engine can vary significantly depending on the domain. But despite this progress, current state-of-the-art speech recognizers perform poorly in domains that require special vocabulary and language models, and under noisy conditions. Furthermore, our previous studies left open the question of whether performance is equivalent for speakers with different accents.

With the advancement of ASR and spoken dialogue systems, 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 than on rural White Californians engaging in sociolinguistic interviews [16]. Thus in [27, 28], we evaluated the performance of popular ASR platforms (Google, Microsoft, Apple, Amazon, IBM, Kaldi) on English speech from populations with different accents. We began with a high-level distinction between General American accents and non-American accents, and then focused on more specific categories of non-American accents including French, Indian, British, and East Asian accents. We reported on a re-analysis of a subset of the ASR outputs examined by [11], including ASR outputs using 2 additional configurations of the Google ASR platform that were not reported in [11], and new annotations for speaker accent. Most ASR systems performed fairly well for General American accents, but all of them did 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 varied approximately from 2% to 12%, with relative differences varying approximately between 16% and 49%.

This drop in performance became even larger when we considered specific categories of non-American accents, e.g., French, British, East Asian, etc.

This performance gap suggests that consumers with non-American English accents may find it considerably harder to take advantage of ASR technology. It is an open research question and an active area of research how to improve ASR systems so that they perform equally well for native and non-native speakers of a language, in our case American English [17, 12, 13, 32, 2, 25, 26, 29]. To improve performance, ASR systems should be trained on more diverse speaker data [9]. This requires more diligent collection of non-American English speaker data.

3 Data

We evaluated Whisper ASR on 6 data sets representing different dialogue domains (5 in total) and types of speaker. Each of our domains involves conversation between a human participant (from the target user population) and one or more virtual characters, except for the domain of the IOTA system (see below) which has conversations between two human participants. The data sets derived from collected interactions of humans and virtual characters include only utterances spoken by human participants, and not by the virtual characters.

- **Amani** [4] is a bargaining character used as a prototype for training soldiers to perform tactical questioning. Speech comes from cadets at the U.S. Military Academy in April 2009, who interacted with Amani as a university course exercise on negotiation techniques. The system maintains dialogue context so the dialogue is not just question answering. Example user utterances while interacting with Amani are: "Hello Amani, how are you today?", "Do you know who did the shooting?", "What do you know about the sniper?", "Do you know where he lives?", "I'll keep this a secret.", "How do you know that?", etc.
- **SGT Blackwell** [18] 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 where he was part of the National Design Triennial exhibition [23]. The museum visitor population includes children and tourists from around the world. The museum exhibit listed a set of about five sample questions, but visitors were free to ask anything they wanted. Example user questions directed at SGT Blackwell are: "What is your favorite color?", "What is your favorite music?", "Who programmed you?", "How come you can understand me?", "Where were you born?", etc.
- **IOTA (Intelligent Operator Training Assistant)** [24] is part of a virtual reality urban combat environment, the Joint Fires and Effects Trainer System (JFETS). Speech for the IOTA domain was collected in 2008 from training sessions in the virtual reality environment at Fort Sill between a human trainee and a human instructor on a variety of missions. We distinguish between Call For Fire (CFF) and Call for Air Support (CAS) missions. Thus the IOTA data set includes both

CFF and CAS relevant conversations whereas the IOTA-FO (IOTA Fires Only) data set only includes CFF relevant conversations. Audio was captured over a simulated radio with reduced sampling rate. Examples of IOTA utterances are: “Roger where do you want hog to look from now that I’m looking at that building, where do you want me to go?”, “From that unit from that intersection go west three units of measure.”, “Okay you mean the y that follows to the southwest?”, “Got a big square field a village on the south side and a lake on the east side.”, “Okay contact on that unit of measure.”, etc. Examples of IOTA-FO utterances are: “Message to observer thunder delay alpha target number alpha bravo zero zero zero six over.”, “Left two hundred over.”, “Repeat target number alpha bravo zero zero zero one one over.”, “Fire for effect over.”, etc.

- **SASO** [30] is a negotiation training prototype in which two virtual characters negotiate with a human “trainee” about moving a medical clinic. Speech was collected at the USC Institute for Creative Technologies during 2006–2009, mostly from visitors and new hires. The system maintains dialogue context so the dialogue is not just question answering. Example user utterances while interacting with SASO are: “I have orders to move this clinic to a camp near the U.S. base.”, “Would you be willing to move downtown?”, “We can build a well for you.”, “We can provide medical supplies.”, “I will protect you from the insurgents.”, “I know this is a big conflict, there’s gonna be a bigger problem if we don’t move the clinic so please cooperate, work with us and we’ll protect you and help you escort all your patients and your supplies to a different location, I’ll have my men move it.”, “It’s not safe here, we can’t protect you.”, etc.
- **SGT Star** [3] is a question-answering character who talks about careers in the Army. Speech collected in the context of the SGT Star system comes from trained handlers who operated SGT Star at job fairs in 2008, presenting to people attending the event. Interaction with SGT Star is typically in the form of independent direct questions, e.g., “Who are you?”, “Is the pay good in the Army?”, “What are the green berets?”, “Just tell us how you can talk.”, “What are special ops?”, “Tell us what the ranger school is like.”, “Have you ever been in combat?”, etc.

The utterances collected from user sessions in the domains described above were transcribed manually to create a separate corpus for each of the domains. We selected utterances from each corpus randomly to create training, development and test sets: development and test sets were each slightly over 10% of the total utterances (dialogue turns) in each corpus, and the remaining utterances were assigned to the training set. In this paper, as well as in [11, 27, 28], we only use the test sets. In Table 1 we report statistics for each domain in terms of word (token) count, number of dialogue turns, and mean turn length (MTL, measured in words). The number of turns is basically the number of audio files for each data set that we use in our evaluation.

For the evaluation of Whisper ASR on different accents we use 2380 utterances from Blackwell. Thus here we use slightly more data than in our previous work [27, 28]. 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 En-

Table 1 Data used in the evaluation: number of words, number of dialogue turns, and mean turn length (MTL); MTL is measured in words.

	#Words	#Turns	MTL
Amani	1855	188	9.9
Blackwell	11520	2500	4.6
IOTA	5441	650	8.4
IOTA-FO	1018	155	6.6
SASO	3483	510	6.8
Star	2137	400	5.3

glish and non-American English accents. We use the term “General American” to encompass the utterances in our data set lacking distinct regional and social characteristics [33, 31]. 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 African American Vernacular English. Next, we segmented the non-American subset further into subcategories of non-American accents, the most common of which in our data set 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.

As reported in [27, 28], 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 non-native 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. Detailed inter-annotator agreement results are reported in [28]. Overall agreement among the annotators was moderate to high, even when including the non-native speaker annotations. Depending on the comparisons Krippendorff’s alpha ranged from 0.672 to 0.901. Krippendorff’s alpha between Annotators 1 and 2 was measured at 0.672 when all 8 distinct categories were considered. Krippendorff’s alpha among all 3 annotators was measured at 0.719 when General American and Northeast American accents were merged into one “American” category, i.e., 7 distinct categories were considered. The results reported below are based on the annotations of Annotator 1.

4 Whisper

Whisper is an ASR system developed by OpenAI [22]. It is trained on 680,000 hours of multilingual and multitask supervised data collected from the web. Whisper models are Transformer sequence-to-sequence models trained on various speech processing tasks, and can be used for speech recognition, speech translation, spoken language identification, and voice activity detection.

The default Whisper implementation is designed to run in offline mode where the ASR has all of the audio available to it at the same time. There are 9 models available from the OpenAI Whisper GitHub[1]. The tiny, base, small, medium, and large models have 39M, 74M, 244M, 769M, and 1550M parameters respectively. There are both English-only and multilingual versions of the tiny, base, small, and medium models. The English-only models are called tiny.en, base.en, small.en, and medium.en respectively. The large model has only a multilingual version. There is also a Whisper large-v2 model available from Hugging Face that we do not use here[2].

We ran Whisper in 2 different ways: (1) using the Python interface where a model is loaded in the beginning and then Whisper processes a series of audio files, one after the other; and (2) using the command line interface where Whisper is called from scratch for each audio file. As expected the Python interface is faster but is not robust. It frequently results in crashes and this happens at random points, i.e., in the same series of audio files it does not always crash while processing the same file. Also, in many cases the Python interface results in worse performance than the command line interface. Problems with the Python interface (e.g., crashes and potential memory leaks) have been reported by several users on web discussion boards. Generally the command line interface is more stable but also considerably slower.

For our experiments we used the following models: tiny, base, small, medium, large, tiny.en, base.en, small.en, and medium.en. The original Whisper ASR is not designed for live speech recognition and the larger the model the slower its performance. As reported on the OpenAI Whisper GitHub, the tiny models (English-only and multilingual) are 32 times faster than the large model, the base models are 16 times faster than the large model, the small models are 6 times faster than the large model, and the medium models are twice faster than the large model. There is active research on equipping Whisper with real-time capabilities such as Whisper-Streaming [19], an implementation of real-time transcription and translation using Whisper models.

5 Experiments

Our main evaluation metric is word error rate (WER). 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

[1] https://github.com/openai/whisper

[2] https://huggingface.co/openai/whisper-large-v2

Table 2 Results in terms of WER (%) using the Python interface for the Amani and SASO data sets. The best result for each data set is shown in bold.

	Amani		SASO	
	English-only	Multilingual	English-only	Multilingual
tiny	11.59	11.98	8.07	10.34
base	10.38	10.77	7.93	7.9
small	9.95	9.95	6.36	6.1
medium	8.79	10.22	6.71	6.53
large	-	**8.35**	-	**5.63**

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.

We first performed experiments using the Python interface on Amani and SASO. Table 2 shows the corresponding WERs for all 9 models. As we can see in Table 2 for Amani the English-only models perform better than their multilingual counterparts. For SASO the English-only version of tiny outperforms the multilingual version of tiny and for the rest of the models the differences between the English-only and multilingual versions are small. The fact that often the English-only version of a model outperforms or performs similar to its multilingual version is also further confirmed by our experiments on evaluating accents (see below). But the large multilingual model performs the best for both Amani and SASO. According to the OpenAI GitHub, for English-only applications the ".en" models tend to perform better than their multilingual counterparts, especially for the tiny.en and base.en models. The difference becomes less significant for the small.en and medium.en models. For this reason, for our evaluation across the 6 data sets we only used the English-only models. Another consideration was speed which is a very important factor when using ASR as part of a dialogue system. As mentioned in section 4, larger Whisper models are slower than smaller ones and that is why we decided not to use the large model. Assessing speed and latency under different configurations, e.g., memory, CPU, GPU, etc., and exploring how Whisper can be used live [19] and in an incremental setting [1, 34] are part of our planned future work.

Table 3 shows results using the command line interface and English-only models (tiny.en, base.en, small.en, medium.en). We can also see the best result from our most recent previous evaluation [11] where in most cases Google ASR performed the best. When Google is not the best ASR from our previous work we report on the best ASR as well as Google. For example, for Blackwell in [11] Apple performed the best followed by Google. In Table 3 we can see that Whisper outperforms the best ASR systems from [11] except for Blackwell. The larger the Whisper model the better its performance except for SASO where small.en outperforms medium.en. For Blackwell, with the medium.en model Whisper performs worse than Apple but better than Google, and with the rest of the models Whisper performs worse than both Apple and Google. Blackwell is the largest of our data sets and contains speech from the general public (including children's speech) and this may explain why

Table 3 Results in terms of WER (%) for English-only models using the command line interface across the 6 data sets. Best previous result refers to the best WER from [11]. The best result for each data set is shown in bold.

	Amani	Blackwell	IOTA	IOTA-FO	SASO	Star
tiny.en	11.54	23.72	33.11	34.66	7.67	18.08
base.en	9.62	19.66	28.27	31.51	6.94	15.23
small.en	9.4	16.99	21.55	22.58	**6.01**	14.29
medium.en	**8.46**	15.2	**19.45**	**19.22**	6.79	**13.76**
Best previous result	11.62 (Google)	**12.66** **(Apple)** 15.91 (Google)	34.9 (Google)	33.51 (Google)	8.53 (Google)	17.64 (Google)

Whisper did not perform as well as for the other data sets. It is interesting that on Blackwell Whisper was outperformed by Apple and in some cases (depending on the Whisper model) by Google even though the versions of these recognizers were older (from 2020). This suggests that Whisper is not as robust as commercial ASR systems on some data sets (see also Tables 4 and 5).

Apart from crashes and instability (see section 4), Whisper also suffered from hallucinations. In several cases it would produce very long sequences of words that had nothing to do with what the speaker said, or it would generate the same word multiple times. We also fed Whisper audio files containing silence only (not included in the data sets we use here) and Whisper would occasionally hallucinate. Whisper hallucinations happened with both the Python and the command line interfaces and seemed random. It is unclear why they happened while processing an audio file, and why one model would produce hallucinations for this audio file but another model would not. It is also unclear why some audio files containing silence resulted in hallucinations and others did not. Note that hallucinations can significantly raise the WER because they typically generate very long sequences of words that the speaker has not uttered (a very large number of insertions).

As mentioned above, to evaluate Whisper ASR models on different accents we used only the Blackwell data. This is the largest of our data sets and also includes a variety of accents from the general public. Table 4 shows results for General American, Regional American, All American, and All Non-American accents. We can also see the best result from our previous evaluation [28] where in most cases Apple, Google, and Microsoft performed the best. The English-only models outperform their multilingual counterparts except for All Non-American where small outperforms small.en. Whisper outperforms all ASR systems from our previous work [28] for Regional American only. For General American, All American, and All Non-American the best ASR is Apple followed by Google and then Microsoft.

Table 5 shows results for Non-American Uncategorized, European Uncategorized, French, British, East Asian, and Indian. We can also see the best result from our previous evaluation [28] where in most cases Apple, Google, and Microsoft performed the best. Again the English-only models outperform their multilingual

Table 4 Results in terms of WER (%) for General American, Regional American, All American, and All Non-American accents. N shows the number of utterances considered per category. Best previous result refers to the best WER from [28]. The best result for each data set is shown in bold.

	General	Regional	All	All
	American	American	American	Non-American
	N=1842	N=105	N=1947	N=433
tiny	25.35	28.21	25.54	27.87
base	20.54	25.18	20.84	25.03
small	14.91	22.14	15.38	18.69
tiny.en	21.13	23.21	21.26	24.04
base.en	17.56	21.07	17.78	19.67
small.en	14.65	**9.46**	14.32	20.38
Best	**10.21**	11.39	**10.95**	**12.52**
previous	**(Apple)**	(Google)	**(Apple)**	**(Apple)**
result	11.24	17.84	11.25	14.61
	(Google)	(Microsoft)	(Google)	(Google)
	15.47	20.3	15.66	18.51
	(Microsoft)	(Amazon)	(Microsoft)	(Microsoft)

Table 5 Results in terms of WER (%) for Non-American uncategorized, European uncategorized, French, British, East Asian, and Indian accents. N shows the number of utterances considered per category. Best previous result refers to the best WER from [28]. The best result for each data set is shown in bold.

ASR	Non-American	European	French	British	East	Indian
	Uncat	Uncat			Asian	
	N=166	N=96	N=42	N=90	N=22	N=17
tiny	29.17	22.39	30.3	30.95	23.23	28.99
base	27.63	21.63	26.67	26.6	15.15	18.84
small	21.6	13.23	16.97	20.97	**14.14**	17.39
tiny.en	27.07	17.81	18.79	28.13	15.15	30.43
base.en	23.7	12.98	16.36	22.51	**14.14**	15.94
small.en	24.4	16.54	13.33	20.46	18.18	20.29
Best	**12.77**	**6.89**	**5.63**	**17.09**	15.15	**12.33**
previous	**(Apple)**	**(Apple)**	**(Apple)**	**(Google)**	(Microsoft)	**(Google)**
result	12.91	11.73	9.38	18.37	19.19	15.07
	(Google)	(Google)	(Microsoft)	(Apple)	(Google)	(Apple)
	19.43	11.73	11.88	23.21	19.19	27.4
	(Microsoft)	(Microsoft)	(Google)	(Microsoft)	(Apple)	(Amazon, IBM)

counterparts except for the small model for Non-American Uncategorized, European Uncategorized, East Asian, and Indian. The multilingual tiny version is also a little better than the English-only tiny model for Indian. Whisper is always worse than the best ASR models from our previous work [28] except for East Asian. For the rest of the data sets Apple, Google, and Microsoft performed the best with the order varying depending on the accent.

6 Conclusion and Future Work

We evaluated multiple off-the-shelf multilingual and English-only OpenAI Whisper models for ASR across 5 diverse dialogue domains (6 data sets) in American English. We also compared the performance of the above models on speech from speakers with General American versus non-American accents. We are not aware of other large-scale evaluations of Whisper for dialogue domains. Overall Whisper models performed very well and for all data sets, except for Blackwell, outperformed other ASR systems from our previous work [11]. On the other hand, Blackwell seems to be one of our most challenging data sets for ASR given that it includes a large number of files from a diverse population, and this is why we selected Blackwell for our evaluation on different accents. It is interesting that on Blackwell, overall Whisper was outperformed by Google, Apple, Microsoft, etc. even though the versions of these recognizers were older (from 2020). This suggests that Whisper is not as robust as commercial ASR systems on some data sets.

Whisper was not always stable and the Python interface would crash quite often and at random points while processing a series of audio files. Using both the Python and the command line interfaces occasionally resulted in hallucinations, which in turn produced higher WERs.

Our evaluation is targeted at non-experts with limited experience in ASR, and we expect it to be useful for dialogue system designers and ASR consumers who are trying to decide whether to switch from a commercial speech recognizer (e.g., Google, Apple, Microsoft) to a Whisper model. Of course for a fairer evaluation we need to compare Whisper with current versions of other ASR systems, which we intend to do in our future work. We ran Whisper offline using the default version. But as part of our future work we want to evaluate Whisper in terms of speed and latency as well as accuracy for live ASR, including incremental speech recognition [1, 34], under different configurations, e.g., memory, CPU, GPU, etc. While performing our experiments and especially for the smaller models and the Python interface Whisper produced results fast on a standard machine just using the CPU (no GPU). As mentioned in section 4, there is active research on equipping Whisper with real-time capabilities such as Whisper-Streaming [19], an implementation of real-time transcription and translation of Whisper models.

Acknowledgements This work was supported by the U.S. Army under Cooperative Agreement Number W911NF-20-2-0053 and Contract Number W911NF-14-D-0005. Statements and opinions expressed and content included are those of the authors and do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Addlesee, A., Yu, Y., Eshghi, A.: A comprehensive evaluation of incremental speech recognition and diarization for Conversational AI. In: Proceedings of the 28th International Con-

ference on Computational Linguistics (COLING), pp. 3492–3503. Barcelona, Spain (Online) (2020)

2. Ahamad, A., Anand, A., Bhargava, P.: 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), pp. 5351–5358. Marseille, France (online) (2020)
3. Artstein, R., Gandhe, S., Gerten, J., Leuski, A., Traum, D.: Semi-formal evaluation of conversational characters. In: O. Grumberg, M. Kaminski, S. Katz, S. Wintner (eds.) Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday, *Lecture Notes in Computer Science*, vol. 5533, pp. 22–35. Springer, Berlin (2009)
4. Artstein, R., Gandhe, S., Rushforth, M., Traum, D.: Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue (SemDial–DiaHolmia). Stockholm, Sweden (2009)
5. Baumann, T., Kennington, C., Hough, J., Schlangen, D.: Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. In: Proceedings of the 7th International Workshop on Spoken Dialogue Systems Technology (IWSDS). Saariselkä, Finland (2016)
6. Broughton, M.: Measuring the accuracy of commercial automated speech recognition systems during conversational speech. In: Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges. Melbourne, Australia (2002)
7. Burger, S., Sloane, Z.A., Yang, J.: Competitive evaluation of commercially available speech recognizers in multiple languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 809–814. Genoa, Italy (2006)
8. Devine, E.G., Gaehde, S.A., Curtis, A.C.: Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. The Journal of American Medical Informatics Association **7**(5), 462–468 (2000)
9. Fukuda, T., Fernandez, R., Rosenberg, A., Thomas, S., Ramabhadran, B., Sorin, A., Kurata, G.: Data augmentation improves recognition of foreign accented speech. In: Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH), pp. 2409–2413. Hyderabad, India (2018)
10. Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., Suendermann-Oeft, D.: Comparing open-source speech recognition toolkits. Tech. rep., Stuttgart, Germany (2014)
11. Georgila, K., Leuski, A., Yanov, V., Traum, D.: Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 6469–6476. Marseille, France (online) (2020)
12. Ghorbani, S., Hansen, J.H.: Leveraging native language information for improved accented speech recognition. In: Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH), pp. 2449–2453. Hyderabad, India (2018)
13. Jain, A., Upreti, M., Jyothi, P.: Improved accented speech recognition using accent embeddings and multi-task learning. In: Proceedings of the 19th Annual Conference of the Speech Communication Association (INTERSPEECH), pp. 2454–2458. Hyderabad, India (2018)
14. Këpuska, V., Bohouta, G.: Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). International Journal of Engineering Research and Application **7**(3), 20–24 (2017)
15. Kim, J.Y., Liu, C., Calvo, R.A., McCabe, K., Taylor, S.C.R., Schuller, B.W., Wu, K.: A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. In: arXiv:1904.12403 (2019)
16. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S.: Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences of the United States of America (PNAS) **117**(14), 7684–7689 (2020)
17. Le, J.T., Best, C.T., Tyler, M.D., Kroos, C.: Effects of non-native dialects on spoken word recognition. In: Proceedings of the 8th Annual Conference of the Speech Communication Association (INTERSPEECH), pp. 1589–1592. Antwerp, Belgium (2007)

18. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proceedings of the 7th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 18–27. Sydney, Australia (2006)
19. Machácek, D., Dabre, R., Bojar, O.: Turning Whisper into real-time transcription system. In: arXiv:2307.14743 (2023)
20. Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., Traum, D.: Which ASR should I choose for my dialogue system? In: Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 394–403. Metz, France (2013)
21. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Waikoloa, Hawaii, USA (2011)
22. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: arXiv:2212.04356 (2022)
23. Robinson, S., Traum, D., Ittycheriah, M., Henderer, J.: 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). Marrakech, Morocco (2008)
24. Roque, A., Georgila, K., Artstein, R., Sagae, K., Traum, D.R.: Natural language processing for joint fire observer training. In: Proceedings of the 27th Army Science Conference. Orlando, Florida, USA (2010)
25. Shibano, T., Zhang, X., Li, M.T., Cho, H., Sullivan, P., Abdul-Mageed, M.: 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 (2021)
26. Sullivan, P., Shibano, T., Abdul-Mageed, M.: Improving automatic speech recognition for non-native English with transfer learning and language model decoding. In: arXiv:2202.05209 (2022)
27. Tadimeti, D., Georgila, K., Traum, D.: How well can an agent understand different accents? In: 5th Widening NLP (WiNLP) Workshop - Co-located with EMNLP. Punta Cana, Dominican Republic (2021)
28. Tadimeti, D., Georgila, K., Traum, D.: Evaluation of off-the-shelf speech recognizers on different accents in a dialogue domain. In: Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC), pp. 6001–6008. Marseille, France (2022)
29. Tong, F., Li, T., Liao, D., Xia, S., Li, S., Hong, Q., Li, L.: The XMUSPEECH system for accented English automatic speech recognition. Applied Sciences **12**(1478) (2022)
30. Traum, D.R., Marsella, S., Gratch, J., Lee, J., Hartholt, A.: Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In: Proceedings of the 8th International Conference on Intelligent Virtual Agents (IVA). Tokyo, Japan (2008)
31. Van Riper, W.R.: General American: An ambiguity. Dialect and Language Variation pp. 123–135 (1986)
32. Viglino, T., Motlicek, P., Cernak, M.: End-to-end accented speech recognition. In: Proceedings of the 20th Annual Conference of the Speech Communication Association (INTERSPEECH), pp. 2140–2144. Graz, Austria (2019)
33. Wells, J.C.: Accents of English, Volume 3: Beyond the British Isles. Cambridge University Press (1982)
34. Whetten, R., Levandovsky, E., Imtiaz, M.T., Kennington, C.: Evaluating automatic speech recognition and natural language understanding in an incremental setting. In: Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue (SemDial). Maribor, Slovenia (2023)
35. Yao, X., Bhutada, P., Georgila, K., Sagae, K., Artstein, R., Traum, D.: Practical evaluation of speech recognizers for virtual human dialogue systems. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC), pp. 1597–1602. Valletta, Malta (2010)