Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6469–6476
Marseille, 11–16 May 2020

Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains

Kallirroi Georgila, Anton Leuski, Volodymyr Yanov, David Traum
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Los Angeles, CA 90094-2536, USA
{kgeorgila, leuski, yanov, traum}@ict.usc.edu

Abstract

We evaluate several publicly available off-the-shelf (commercial and research) automatic speech recognition (ASR) systems across diverse dialogue domains (in US-English). Our evaluation is aimed at non-experts with limited experience in speech recognition. Our goal is not only to compare a variety of ASR systems on several diverse data sets but also to measure how much ASR technology has advanced since our previous large-scale evaluations on the same data sets. Our results show that the performance of each speech recognizer can vary significantly depending on the domain. Furthermore, despite major recent progress in ASR technology, current state-of-the-art speech recognizers perform poorly in domains that require special vocabulary and language models, and under noisy conditions. We expect that our evaluation will prove useful to ASR consumers and dialogue system designers.

Keywords: off-the-shelf speech recognizers, dialogue systems, diverse dialogue domains

1. Introduction

In this paper we evaluate several publicly available off-the-shelf (commercial and research) automatic speech recognition (ASR) systems, using data collected from deployed spoken dialogue systems as well as from human-human conversations in 6 domains (in US-English).

A natural language dialogue system usually takes human speech as its input, thus a speech recognizer resides at the very front-end of such a system. Other natural language processing units highly rely on the output of a speech recognizer. Thus a speech recognizer directly affects the overall system performance. An ASR system mainly uses two models: an acoustic model responsible for modelling the sounds that make up words, and a language model responsible for modelling word sequences. The two models work together to find the best hypotheses (word sequences) corresponding to a given speech signal.

Every dialogue system has a target user population. For example, a conversational assistant, such as Amazon Alexa, Apple Siri, Google Home, or Microsoft Cortana, handles various questions in a broad domain while a dialogue system developed for military purposes must understand military terms (see below). Thus there are many factors to consider in selecting an ASR system for a particular application, among them:

- The domain and vocabulary that the speech recognizer is expected to handle.
- The acoustic environment in which the speech recognizer operates.
- The time it takes for a speech recognizer to generate an output. There is often a trade-off between the quality of the ASR output and the time it takes to generate that output; real-time dialogue systems may be willing to accept a somewhat degraded output in return for lower latencies.
- Whether the speech recognizer can generate incremental outputs or waits until the speaker has finished speaking to generate a complete output.
- Whether the speech recognizer runs on the cloud or can be used on a device without Internet connection. This can be a major issue when there are data privacy concerns.
- The procedure for adapting the speech recognizer to a particular domain by building domain-specific acoustic and/or language models.
- The possibility for training on individual speakers, and the amount of available user-specific training data.

The evaluation described in this paper is targeted to ASR consumers and potential consumers with limited experience in ASR. We use state-of-the-art ASR systems that have been developed both in industry and academia, and our focus is on employing out-of-the-box acoustic and language models, i.e., we do not train domain-specific models.

This is our third large-scale ASR evaluation using corpora from a variety of domains (Yao et al., 2010; Morbini et al., 2013). Compared to our previous evaluations, we see a large improvement in ASR performance, which illustrates the significant progress that has recently been made in ASR technology, especially with the use of deep learning techniques. However, there are domains where interactions take place under noisy conditions and that require special vocabulary and language models. In these domains we will see that current state-of-the-art speech recognizers perform poorly. Furthermore, the performance of a specific ASR system can vary significantly depending on the domain.

The remainder of the paper describes related work, the data used, the ASR engines, the results of our evaluation, as well as discussion on how much ASR technology has advanced since our previous evaluations.

2. Related Work

In one of the earliest studies on ASR evaluation, Devine et al. (2000) compared 3 commercial ASR software packages: 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 data that the ASR systems were tested on were medical progress notes and discharge summaries drawn from actual records and dictated by 12 physicians after minimal training with each software package. The IBM system performed the best.

In a recent study, also in a medical domain, Kim et al. (2019) tested 5 ASR platforms in terms of transcription quality. In particular, they measured the performance of Google Cloud, IBM Watson, Microsoft Azure, Trint, and YouTube on data collected from the interaction of 12 medical students with 2 simulated patients (12 dyadic medical teleconsultations in total). Note that simulated patients are human actors trained to act as patients in a medical situation. Kim et al. (2019) did pairwise comparisons between each of the 5 ASR systems outputs and manual transcriptions. 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 (2002) evaluated 2 commercial speech recognizers 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. (2006) evaluated 3 commercial desktop dictation ASR engines in 8 languages (US-English, UK-English, Iberian Spanish, French, German, Japanese, Simplified Chinese, and Traditional Chinese). They found the performance of the ASR systems to be 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. (2014) 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. (2014) 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 (2017) 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. (2016) 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.

Our first large-scale ASR evaluation was done in 2010 (Yao et al., 2010). 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 6 data sets.

Our second large-scale ASR evaluation was done in 2013 (Morbini et al., 2013). This evaluation included 2 research platforms, i.e., CMU PocketSphinx, Otosense-Kaldi (a system developed at USC based on the research toolkit Kaldi (Povey et al., 2011)); 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.

Both of our evaluations (Yao et al., 2010; Morbini et al., 2013) used much more diverse data sets and domains than previous work. We also employed a larger variety of ASR systems (both commercial and research) compared to previous work. In this paper we use mostly the same data sets as in our previous evaluations but our focus is on employing out-of-the-box acoustic and language models, and of course we use the latest state-of-the-art publicly available speech recognizers (both commercial and research).

3. Data

We evaluated the ASR systems on 7 data sets representing different dialogue domains (6 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 Section 3.5.), 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.

In the remainder of this section, we describe each of the domains used in our evaluation, including brief descriptions of the dialogue systems with which the utterances were collected.

3.1. SGT Star

SGT Star (Artstein et al., 2009a) 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. The following utterances illustrate questions asked by handlers using speech recognition, and SGT Star's answers:

Handler Who are you?

SGT Star Who, me? Oh, I'm Sergeant Star! I'm a virtual character. Maybe you've seen me before on the web at GoArmy dot com. But there, I'm stuck looking like this. But now with a brilliant blend of pixels, polygons and programming, coupled with real time animation and spoken word interface, I can do my very best to answer your questions on Army careers. And my lips move when I do it.

Handler Is the pay good in the Army?

SGT Star The Army is a career and like most jobs, you start at the bottom and work your way up. And when you throw in the allowances for housing, meals, uniforms, medical benefits, we're very competitive with the civilian world.

3.2. Amani

Amani (Artstein et al., 2009b) 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 following utterances illustrate a segment of a dialogue between a cadet and Amani:

Cadet Do you know who did the shooting?

Amani So you want to talk about the man who did the shooting? I have heard that he is called Saif.

Cadet Do you know where he lives?

Amani You are asking for a lot of information. If you expect me to tell you that you must offer me secrecy.

Cadet I will keep this a secret.

Amani You have assured me of secrecy. I believe he hides on the second floor of Assad's shop. I know he is in there but when I go in the shop I don't see him. And I have no idea where he is now.

3.3. SGT Blackwell

SGT Blackwell (Leuski et al., 2006b) 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, like SGT Star. However, the questions collected from sessions with SGT Blackwell come from the general public, and not from trained handlers as was the case with the questions collected from sessions with SGT Star. 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.

3.4. SASO

The SASO system (Traum et al., 2008) 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 following utterances illustrate a segment of a dialogue between a trainee and two virtual characters (Doctor Perez and Elder Al Hassan) in the SASO system:

Trainee I have orders to move this clinic to a camp near the U.S. base.

Elder Al Hassan We have many matters to attend to.

Trainee I understand, but it is imperative that we move the clinic out of this area.

Elder Al Hassan This town needs a clinic.

Doctor Perez We can't take sides.

Trainee Would you be willing to move downtown?

Elder Al Hassan We would need to improve water access in the downtown area, captain.

Trainee We can dig a well for you.

Doctor Perez Captain, we need medical supplies in order to run the clinic downtown.

3.5. IOTA

IOTA (Intelligent Operator Training Assistant) (Roque et al., 2010) 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 utterances from a complex mission spoken by a trainee and an instructor are shown below:

Trainee Roger where do you want hog to look from now that I'm looking at that building, where do you want me to go?

Instructor Follow the y to the south.

Trainee Okay you mean the y that follows to the southwest?

Instructor Affirmative.

Trainee Roger contact on that east west road.

Instructor From that unit from that intersection go west three units of measure.

	#Words	#Turns	MTL
Star	2137	400	5.3
Amani	1855	188	9.9
Blackwell	11520	2500	4.6
SASO	3483	510	6.8
IOTA	5441	650	8.4
IOTA-FO	1018	155	6.6
Twins	8279	2017	4.1

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

3.6. The Twins

The Twins are two life-size virtual characters who serve as guides at the Museum of Science in Boston (Swartout et al., 2010). The characters promote interest in Science, Technology, Engineering and Mathematics (STEM) in children between the ages of 7 and 14. They are question answering characters, but unlike SGTs Blackwell and Star, the response is a whole dialogue sequence, potentially involving interchange from both characters, rather than a single character turn.

There are two types of users for the Twins: demonstrators, who are museum staff members, using head-mounted microphones, and museum visitors, who use a Shure 522 table-top mounted microphone (Traum et al., 2012). Speech was collected at the Museum of Science in Boston from 2009 to 2011. More on analysis of the museum data can be found in (Aggarwal et al., 2012). The following utterances illustrate a segment of a dialogue between a museum visitor and the Twins:

Museum visitor Hello.

Ada Hey there!

Grace How can we help you?

Museum visitor Who are you?

Grace Could you say that again?

User Who are you named after?

Ada Our namesakes are Ada Lovelace and Grace Hopper.

Grace Yeah, they may be old fashioned names. But both are from the most modern computing women of their time.

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 we do not train our own acoustic or language models so we only used 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).

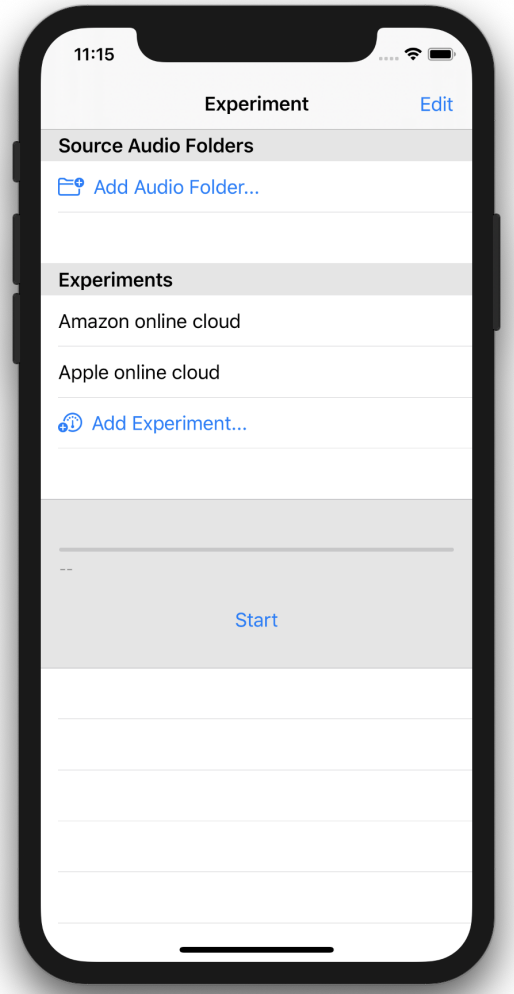

Figure 1: A screenshot of the ASR testing app.

4. 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. Below we provide more details about the setup of each of these platforms used in our experiments.

We were looking for a single common platform where we could test all of the commercial ASR systems. While Amazon, Google, IBM, and Microsoft provide SDKs in several different languages and support a number of different platforms, Apple only supplies ASR for iOS and macOS. However, their support for macOS is limited to the cloud-based ASR at the moment. Google, IBM, and Microsoft do not officially support macOS.

Considering these limitations, we ended up developing a test application for iPhone. Figure 1 shows a screenshot of the app. Here a user selects a set of directories with audio files and configures a collection of ASR systems to apply to the files. The app sends each audio file through each of the selected systems, collects the transcripts, and stores them

into a single JSON file.

Most of our testing of commercial ASR systems was done in online mode. 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. 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.

In contrast to the commercial speech recognition platforms we conducted our Kaldi experiments on a local desktop machine.

4.1. Amazon

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

4.2. 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. The cloud access is free, however Apple limits 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 has no limitations. In this study we used both the cloud-based ASR and the on-device ASR running on iPhone XS.

4.3. 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). The service is 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 and default models.

4.4. 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. The first 500 minutes per month are free and between $0.02 and $0.01 per minute afterwards depending on the usage.

4.5. 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.

4.6. 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. The service is free for the first 300 minutes each month and $0.016 per minute afterwards. We ran the system in both offline and online modes.

4.7. Summary

Table 2 provides a summary of each one of the configurations that we used. Kaldi is only available to run on a device. The rest of the ASR systems run on the cloud, except for the Apple one which also runs on a device.

5. Results

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 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 can be formulated as:

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

[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

[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 default	cloud	online	default
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.

ASR	Blackwell	STAR	Twins	Amani	SASO	IOTA	IOTA-FO
Amazon cloud online	21.43	22.54	20.47	21.86	17.90	51.07	44.01
Apple device online	16.02	24.44	16.28	15.69	12.17	52.22	47.90
Apple cloud online	**12.66**	21.40	13.23	13.27	11.09	47.45	46.22
Google cloud online default	16.58	22.02	12.58	13.88	12.72	45.53	45.06
Google cloud online video	15.91	**17.64**	**9.50**	**11.62**	**8.53**	**34.90**	**33.51**
IBM cloud online	31.72	19.64	24.58	13.11	12.81	41.21	35.92
Kaldi device offline ASpIRE	31.01	25.91	30.74	20.21	18.70	43.36	44.12
Kaldi device online ASpIRE	37.78	26.20	32.67	22.28	20.34	47.77	49.37
Kaldi device offline LibriSpeech	47.99	40.10	59.74	25.77	24.13	73.38	77.10
Kaldi device online LibriSpeech	51.60	41.20	59.40	25.49	25.02	75.78	80.67
Microsoft cloud offline	18.17	23.54	21.24	29.19	15.98	44.57	40.44
Microsoft cloud online	18.93	22.35	24.42	22.30	16.09	45.33	42.96

Table 3: Results in terms of WER (%).

Transcription: that's affirmative focus on the lake ASR output: affirmative focus on the
Transcription: uh contact with that target let me go ahead and talk to my wingman ASR output: contact with that target let me go in trouble
Transcription: hawk two two in from the north tally target ASR output: okay a tutu is the north
Transcription: confirm that plume is north of the target ASR output: term that pumas north of the target
Transcription: thunder four zero bandit four two repeat over ASR output: butterfly zero bandit for to repeat
Transcription: good copy that is your target ASR output: good copy that is your talk
Transcription: one platoon neutralized estimate zero one two casualties out ASR output: bumper to neutralize estimate zero one two casualties out
Transcription: message to observer target correction alpha illumination target number alpha bravo ASR output: let's deliver targeted correction mess to observer alpha illumination part number alpha bravo

Table 4: Examples of errors generated by the Google cloud online video ASR system on the IOTA data set (best performing ASR system on this data set).

Table 3 summarizes our results. Several conclusions can be drawn from the results. First, there are a lot of errors in many domains. This underscores the point that ASR for conversational speech is still a challenging task and further work is needed on ASR performance and NLU and dialogue techniques to cope with high error rates (Leuski et al., 2006a). Second, there are large differences in the recognition rates for the different domains. Some of these

	Current evaluation		Previous evaluation	
	WER	Best ASR	WER	Best ASR
Blackwell	12.66	Apple cloud online	18.00	Google from 2013
Star	17.64	Google cloud online video	21.70	AT&T from 2013
Twins	9.50	Google cloud online video	18.70	Otosense-Kaldi from 2013
Amani	11.62	Google cloud online video	23.80	Google from 2013
SASO	8.53	Google cloud online video	16.30	AT&T from 2013
IOTA	34.90	Google cloud online video	39.00	HDecode from 2010

Table 5: Comparison with previous evaluations on the same data sets in terms of WER (%) – best current result and best previous result for each data set.

differences may be an artifact of the size of the collected data set, but other aspects concern the domain itself, e.g., mean turn length, size of vocabulary, and how specialized the vocabulary is. Third, no one recognizer dominates on all data sets.

Overall Google cloud online video performs the best except for the Blackwell domain where Apple cloud online has the lowest WER. 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 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.

All ASR systems performed poorly on the IOTA and IOTA-FO data sets because of the specialized military language and terms. The best performing ASR system on these data sets was Google cloud online video with a WER of 34.90% and 33.51% respectively, which is considered quite high. Clearly, for the IOTA domain there is a strong need for customized models. Table 4 shows examples of errors generated by the Google cloud online video ASR system on the IOTA data set.

Compared to our previous evaluations (Yao et al., 2010; Morbini et al., 2013), WERs are now significantly lower. Table 5 shows for each data set the best current result and the best previous result (either from the 2010 evaluation or the 2013 evaluation). All these results are with out-of-the-box models except for Otosense-Kaldi on the Twins data set and HDecode on the IOTA data set which used domain-specific acoustic and language models. Also, note that our best previous result for Twins using out-of-the-box models was with Google (20.60%). Overall absolute improvements in WER values range from about 4-5% for Blackwell, Star, and IOTA to about 8-9% for Twins and SASO, and about 12% for Amani. The relative improvements in WER values are even more impressive: about 50% for Twins, Amani, and SASO, about 30% for Blackwell, about 20% for Star, and about 10% for IOTA. This illustrates the fact that there has been major progress in ASR technology in recent years.

6. Conclusion

We evaluated several publicly available off-the-shelf (commercial and research) ASR systems across diverse dialogue domains (in US-English). Our evaluation is aimed at non-experts with limited experience in speech recognition. For this reason, we did not train domain-specific acoustic or language models.

Our results show that the performance of each speech recognizer can vary significantly depending on the domain. Comparison with our previous evaluations from 2010 and 2013 on the same data sets shows that there has been major progress in ASR technology in the last few years, especially with the use of deep learning techniques. 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. We expect that our evaluation will prove useful to ASR consumers and dialogue system designers.

For future work, we plan to train domain-specific language models, interpolate them with general-purpose language models, and see whether this leads to lower WERs, especially for the IOTA domain where clearly there is a strong need for domain-specific models. We will also perform tests on additional data sets and experiment with different configurations of the ASR systems.

7. Acknowledgments

The work depicted here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

8. Bibliographical References

Aggarwal, P., Artstein, R., Gerten, J., Katsamanis, A., Narayanan, S., Nazarian, A., and Traum, D. (2012). The Twins corpus of museum visitor questions. In *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 2355–2361, Istanbul, Turkey.

Artstein, R., Gandhe, S., Gerten, J., Leuski, A., and Traum, D. (2009a). Semi-formal evaluation of conversational

characters. In Orna Grumberg, et al., editors, *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *Lecture Notes in Computer Science*, pages 22–35. Springer, Berlin.

Artstein, R., Gandhe, S., Rushforth, M., and Traum, D. (2009b). Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In *Proc. of the 13th Workshop on the Semantics and Pragmatics of Dialogue (SemDial–DiaHolmia)*, Stockholm, Sweden.

Baumann, T., Kennington, C., Hough, J., and Schlangen, D. (2016). Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. In *Proc. of the 7th International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, Saariselkä, Finland.

Broughton, M. (2002). 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.

Burger, S., Sloane, Z. A., and Yang, J. (2006). Competitive evaluation of commercially available speech recognizers in multiple languages. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 809–814, Genoa, Italy.

Devine, E. G., Gaehde, S. A., and Curtis, A. C. (2000). 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.

Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., and Suendermann-Oeft, D. (2014). Comparing open-source speech recognition toolkits. Technical report, Stuttgart, Germany.

Këpuska, V. and Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *International Journal of Engineering Research and Application*, 7(3):20–24.

Kim, J. Y., Liu, C., Calvo, R. A., McCabe, K., Taylor, S. C. R., Schuller, B. W., and Wu, K. (2019). A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. In *Preprint arXiv:1904.12403*.

Leuski, A., Kennedy, B., Patel, R., and Traum, D. (2006a). Asking questions to limited domain virtual characters: How good does speech recognition have to be? In *Proc. of the 25th Army Science Conference*, Orlando, Florida, USA.

Leuski, A., Patel, R., Traum, D., and Kennedy, B. (2006b). Building effective question answering characters. In *Proc. 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 *Proc. 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 *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Roque, A., Georgila, K., Artstein, R., Sagae, K., and Traum, D. R. (2010). Natural language processing for joint fire observer training. In *Proc. of the 27th Army Science Conference*, Orlando, Florida, USA.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J., Gerten, J., Chu, S., and White, K. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In *Proc. of the 10th International Conference on Intelligent Virtual Agents (IVA)*, Philadelphia, Pennsylvania, USA.

Traum, D. R., Marsella, S., Gratch, J., Lee, J., and Hartholt, A. (2008). Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proc. of the 8th International Conference on Intelligent Virtual Agents (IVA)*, Tokyo, Japan.

Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., and Swartout, W. (2012). Ada and Grace: Direct interaction with museum visitors. In *Proc. of the 12th International Conference on Intelligent Virtual Agents (IVA)*, Santa Cruz, California, USA.

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 *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 1597–1602, Valletta, Malta.