Voice AI Whitepaper

https://aimultiple.com

Executive Summary

Voice AI brings together voice understanding (voice-to-text) and conversational AI (natural language understanding and generation) capabilities to help companies serve current or potential customers over voice channels such as phone or instant messenger calls.

Voice AI applications allow companies to reduce waiting time for customers or potential while reducing the cost to serve them. However, voice bots can reduce customer satisfaction if they fail to understand user intent. Effective voice bots should at least be intelligent enough to understand when they don't understand user intent or when they can not serve the user effectively. This would allow them to pass the conversation to a human operator.

Our recommendation to companies is to identify customer service domains where they can rapidly test voice AI solutions while keeping track of important metrics like NPS. Your assessment of voice technology from 2018 can be significantly outdated with the current technology thanks to recent advances in the field. By rapidly testing vendors and adopting successful solutions, companies can boost customer satisfaction while reducing customer service costs: the holy grail of every business!

The articles included in this whitepaper should allow you to identify voice AI use cases in your business, key criteria for selecting a voice AI solution and start your search with a good understanding of the voice AI ecosystem.

Table of Contents

Voice recognition

Artificially intelligent machines are becoming smarter every day. Deep learning and machine learning techniques enable machines to perform many tasks at the human level. In some cases, they even surpass human abilities. Machine intelligence can analyze big data faster and more accurately than a human possibly can. Even though they cannot think yet, they see, sometimes better than humans (read our computer vision and machine vision articles), they can speak, and they are also good listeners. Known as “automatic speech recognition” (ASR), “computer speech recognition”, or just “speech to text” (STT) enables computers to understand spoken human language.

Speech recognition and speaker recognition are different terms. While speech recognition is to understand what is told, speaker recognition is to know the speaker instead of understanding the context of the speech that can be used for security measures. These two terms are confusing and voice recognition is often used for both.

History

In the 1950s, a system for single-speaker digit recognition developed by three Bell Labs researchers had the capacity of ten words.

Graduated from Stanford university, Raj Reddy tried to develop a system that can recognize continuous speech unlike the previous system that requires pauses between each word. Raj designed the system to enable spoken commands for chess.

Soviet researchers developed a dynamic time warping algorithm that has 200-word vocabulary around the same era. The speech was processed by dividing it into short frames, and The DTW algorithm signal processed each frame as a single unit.

In 1971, with the participation of BBN, IBM, Carnegie Mellon and Stanford Research Institute DARPA funded a program for speech recognition with the goal of a minimum vocabulary size of 1000 words.

In 1980s, a voice activated typewriter called Tangora was created by IBM. Tangora could handle a 20,000-word vocabulary.

The developments are correlated with the hardware. In 1970s, the best computer had 4 MB RAM and it took almost two hours to decode 30 second speech with these computers. As hardware problems were solved, researches could tackle harder problems such as larger vocabularies, speaker independence, noisy environments and conversational speech.

How does voice recognition technology work?

Computers need digitized data to process and analyze. Analog-to-digital converter (ADC) translates the vibrations and analog waves into digital data that computers can analyze. Voice recognition software separated the signal into small segments to match these segments to known phonemes in the related language. The program compares the phonemes words in its built-in dictionary to determine what can be told.

Models

Nowadays, voice recognition systems are using statistical modeling systems that use complex probability and mathematical functions to determine the most likely outcome. The most common methods that are used for voice recognition are Hidden Markov Model and neural networks.

HMM enables information like acoustics, language and syntax in a unified probabilistic model. In this model, programs score each phoneme that will come after another to predict the best next possible phoneme. This process gets harder in phrases and sentences because the system must understand the start and end of the words. When speech gets faster, voice recognition programs can misunderstand. To give an example here is breakdown of the two similar phrases:

- r eh k ao g n ay z s p iy ch
- “recognize speech”
- r eh k ay n ay s b iy ch
- “wreck a nice beach”

Accuracy

None of the voice recognition systems is perfect. Performance of the voice recognition systems is evaluated based on two criteria; speed and accuracy. Accuracy is rated with word error rate (WER). Accuracy may decrease depending on some factors. Some of them are hardware problems like low-quality sound cards, low-quality microphones, or insufficient processor.

Speech recognition requires clean audio and processing power to run the statistical models.

Other than hardware problems, overlapping speech can reduce the accuracy. In the meetings with multiple speakers that constantly interrupts their talks, voice recognition programs often fail to perform efficiently.

Homonyms are also a challenge for voice recognition systems. The words like “there – their”, “air – hair” and “be – bee” that are pronounced similarly but have different meanings are hard to understand from only the sound. To increase the accuracy, extensive training and statistical methods are used.

Applications

Everybody knows Siri, the smart assistant of iPhone users. Siri is the most common example of voice recognition application. The other assistants like Microsoft’s Cortana or Amazon’s Alexa are the best examples of voice recognition-powered programs. Or maybe some of you can recall Jarvis from Ironman.

I guess many of you did use Google's voice to learn the true pronunciation of a word from Google translate. In that case, natural language processing is also used with voice recognition.

YouTube also uses speech recognition to automatically generate subtitles for the videos. When you upload a video that includes speeches or talks, YouTube detects it and provides a transcription. You can also have the minute-by-minute text of the transcribed speech.

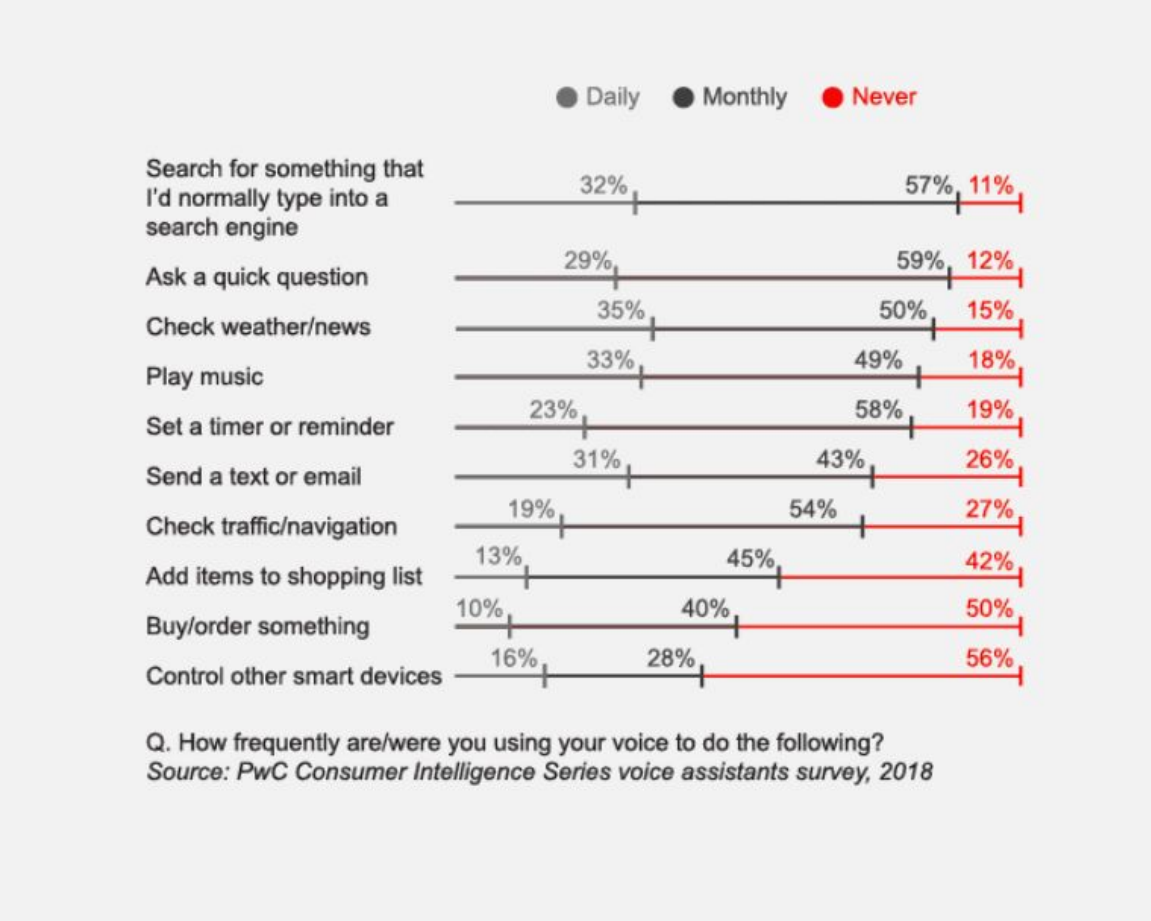
There are many applications where voice recognition is implemented. Even in health care, voice recognition is used. Doctors can determine a person’s mental state whether he/she is depressed or suicidal by analyzing his/her voice.

Voice recognition includes, but not limited to these applications. We will explain most common applications in the next chapter.

Voice recognition applications

Common applications

Voice recognition is a maturing technology and users seem to trust it for the most basic functionality like search or playing music. User adoption of voice interfaces is still low for applications with more significant implications like buying things or controlling smart devices.



SOURCE: PWC

Voice Search

This is the most common use of speech recognition. In 2019, reports estimate that 112 million people in the US will use a voice assistant at least monthly, up 10% from last year. Another study also reveals, approximately 7 out of 10 consumers (71%) prefer to use voice searches to conduct a query over the traditional method of typing. Thanks to applications such as Siri and Google voice search, voice interfaces have become commonly used.

Voice to text

Thanks to speech recognition, you don't need to type emails, reports, and other documents. For instance, you can use these voice typing and voice command features in Google Docs if you are using the Google Chrome browser.

Voice commands to smart home devices

Smart home applications are mostly designed to take a certain action after the user gives voice commands. Smart home devices are widely used speech recognition application, specifically when considering these:

- the total number of smart devices supported by voice assistants tripled between 2018 and 2019
- 30% of voice assistant users state smart home devices as their primary reason(s) for investing in an Amazon Echo or Google Home.

Business function applications

Customer service

This is one of the most important AI applications in customer service. Speech recognition is an effective call center service solution that is available 24/7 at a fraction of the cost of a team of customer service representatives. It transcribes thousands of phone calls between customers and agents to identify common call patterns and issues.

Voice Biometrics for Security

Voice biometrics use a person's voice as a unique identifying biological characteristic in order to authenticate them. Speech recognition can also be used for voice authentication to replace processes where a user has to display her personal information to authenticate herself.

Voice biometrics improves the overall customer experience since it eliminates customer frustration due to cumbersome login processes as well as lost and stolen credentials.

Industry applications

Automotive

In-car speech recognition systems have become a standard feature for most new vehicles. These systems aim to remove the distraction of looking down at your mobile phone while you drive. Thanks to these systems, drivers can use simple voice commands to initiate phone calls, select radio stations or play music.

Academic

80% of sighted children’s learning is through vision and their primary motivator to explore the environment around them. Speech recognition has the potential to minimize the disadvantages of students who are blind or have low vision.

There are also language learning tools such as Duolingo that use speech recognition to evaluate a user's language pronunciation.

Media / Marketing

Tools such as dictation software can enable people to write around 3000-4000 words of content including articles, speeches, books, memos, emails in 30 minutes if they are familiar with the topic. Though these tools still don’t provide 100% accurate results, they are beneficial for first drafts.

Healthcare

During patient examinations, doctors shouldn’t worry about taking notes of patients’ symptoms. Medical transcription software uses speech recognition to capture patient diagnosis notes. Thanks to this technology, doctors can shorten the average appointment which enables doctors to see more patients during their working hours.

Voice bots

What is a voice bot?

Voice bots, also called voice-enabled chatbots, are AI-based software that take voice commands and reply by voice. They enable users to communicate faster compared to text based bots. Popular examples of voice bots Apple's Siri, Amazon Alexa and Google Assistant. There are two types of voice bots:

- Hybrid model: Voice and text controlled bots.
- Voice-only bots: Only Voice-controlled bots.

Like chatbots, voice bots are able to not just recognize what the user says but to understand the customer's intent and have two-way communication to solve the users' problems. In addition to the technologies used in text based bots, voice bots also rely on transcription to first get the user's commands in text form. They also rely on text-to-speech conversion to talk to users.

How do voice bots work?

Voice bots work like text chatbots.. An additional voice recognition step is required in voice bots compared to chatbots.

The steps can be summarized roughly:

1. Voice input is taken from the user with a device like a mobile phone or computer with a microphone.
2. This input is sent to the cloud in order to decode the message and understand the user intent.
3. The audio message is converted to text and the natural language processing models analyze the users' requests.
4. The AI-based engines search for the most suitable answers or actions and create a response.
5. The answer is converted audio and shared with the user.

Why are voice bots important?

Using voice bots provides a better, more natural user interface. Companies can use voice bots to reduce call center costs and create more user friendly products

The increase in mobile users results in an increase in the voice bot demand

Mobile devices provide ease of use in terms of voice applications. It is faster to describe any problem by talking than typing. The progression of speech and voice recognition technologies are supported by tech giants including Google, Apple, Microsoft and Amazon, as well as Baidu, Xiaomi and Alibaba.

Voice-bot adds value to contact centers by reducing queue time and therefore improving customer satisfaction.

The problem-solving process of users with call centers takes less time than the waiting period. By working 24/7 and with the ability to scale up according to demand, voice-bots eliminate wait times.

The market size is increasing as the demand for chatbots and voice bots

IBM points out that businesses spend $1.3 trillion on 265 billion customer service calls each year. The process can be automated partially by using voice bots.

Why choose voice bots over chatbots

There are two main reasons why voice bots are preferred to chatbots.

Speaking is a faster way to explain a problem than typing

Especially for older people, typing is slower than talking in describing a problem. On the other hand, according to PWC research, younger people tend to use voice assistants. It shows that voice bots are more useful for both younger and older people in terms of ease of use and faster communication.

People tend to prefer a human-like interaction to tell their problems

Voice bots provide smoother, more human-like interactions.

Recent advances in voice technologies

Many tech giants invested in voice technologies in recent years. A disadvantage of voice bot over textual chatbots is the speech recognition time. With the advances in both natural language understanding and speech recognition technologies, launching a voice bot is not significantly more challenging than launching a text based chatbot.

How to choose a voice bot platform?

A recent Gartner survey states that the market for conversational AI, chatbots and virtual assistants includes as many as 1,000 to 1,500 vendors worldwide. Therefore choosing the right vendor is crucial. There are four different points to consider while choosing a voice bot provider.

NLU capabilities

The NLU infrastructure sets the boundaries of what a voice bot can do. At this point, there may be a tradeoff between price and NLU capabilities. The optimal NLU skills should be chosen by considering the requirements of the use case.

Deployment

The training process of AI used by the voice bot must be completed in order to minimize the time required for full integration. Installation of the voice bot into the system and without being fully trained can cause problems such as user satisfaction and security. A trained AI can reduce time of deployment.

Use Cases

It is necessary to determine exactly what purpose voice bot to use. It is necessary to determine exactly what purpose voice bot to use. For example, an IVR bot should not have to have complex NLU. Taking different types of work over the same voice-bot increases the likelihood of errors. Therefore, it may be more suitable to choose an application-specific bot.

Pricing

Voice bots have different payment models, just like chatbots. Options such as monthly plan, pay per use, pay per performance may be available. At this point, a payment plan should be selected considering how much the boat will be used. However, in most cases performance-oriented payment plans may be a viable option.

Ability to manage bias

Training datasets can contain more data about certain gender, age, accent or other demographic attributes. This may make voice bots prone to misunderstand

your users when they do not share the common characteristics of your training database. For example, Zaion.ai team underlines that they are developing algorithms to mitigate biases and ensure that no speech is left misunderstood.

Conversational AI

Businesses are adopting conversational technologies to improve experiences for both internal employees and external customers. Conversational platforms are one of the most common uses for AI applications. Businesses aim to improve customer experience and also reduce costs, by integrating the right conversational AI technology.

What is conversational AI?

Conversational artificial intelligence is the technology that enables automatic messaging and conversation between computers and humans. It enables companies to launch chatbots and virtual assistants.

Conversational AI programs can communicate like a human by understanding the purpose in speech or text and imitating human speech. The ultimate goal of conversational AI is to become indistinguishable whether it is a computer or a human being. Designing the flows that sound natural is an important constraint of a conversational AI.

The benefit of conversational AI technology is that it offers customers a direct channel through which they can communicate naturally. The benefit of conversational AI technology is that it offers customers a direct channel through which they can communicate naturally. customers can ask questions through text or voice in order to find answers to their concerns.

How is AI used in conversational platforms?

The simplest example of conversational platforms are structures that send certain outputs to specific inputs. However, thanks to machine learning, conversational platforms can handle a wider range of queries. Additionally, conversational AI systems can consider the context (i.e. the rest of the conversation) while determining the users' intent and the response.

Natural Language Processing (NLP)

NLP is a sub-branch of artificial intelligence that allows you to break down, understand, process and determine the required action. NLP is the engine that performs tasks such as dialog control and task prediction.

- **Dialogue control:** According to the general flow of speech, the perception of conversational AI is shaped and dialogue control modules are used to control pragmatic adaptations in order to make the conversation natural.
- **Task prediction:** Speech flow gives an idea about the intention of the user(to buy something) is estimated and his/her action is recorded.

Natural Language Understanding (NLU)

NLU is a subcategory of NLP that analyzes sentence structures in text and speak formats. NLU enables computers to interpret the meaning in intent with common human errors like mispronunciations or transposed letters. NLU engines are fed with big data and they need verification. Technology giants like Google improve these engines by using their data.

Natural Language Generation (NLG)

Another sub category of NLP, this technology enables the response generation to the user. In order for the speech to be persuasive and fluent, natural answers must be produced to the user.

What are the things to pay attention to while choosing conversational AI solutions?

Conversational AI has to take many factors into consideration in order for the person to understand what they want to tell. At this point, artificial intelligence should go a little further and behave intuitively. In addition, the platform must be secure to protect personal data.

Performance

The user's intent must be understood, no matter how complex the sentence. The language supports must have a wide range. A multinational company should support many languages.

For example, understanding spoken word in SouthEast Asia is a challenge since

- There are more than 1,000 spoken languages in SEA
- In China, there are more than 20 dialects

Security and Privacy

The platform must provide the security of customers' personal information and security of the personal data. Let's assume a bank has a conversational ai platform. The level of data security of the customer is directly related to the reputation of the company.

Integration

The conversational AI platform must be integrated well into existing applications or systems. The testing frameworks and techniques are used to make sure everything is going well.

User Interface

Providing a seamless platform for users will enable them to communicate with conversational platforms more often.

Conversational AI development best practices

Identify your target audience and understand their needs

Improving customer experience is the first reason why most companies are deploying a chatbot. You should understand who your customers are. This will enable you to create relevant conversation flows including FAQs and special offers.

Restrain your ambition and set realistic goals

The scope of your conversational AI is crucial. Save for 2 bots, all bots in the list are laser-focused. For example, Wordsmith takes in structured data to prepare reports, visabot prepares immigration forms.

Even Facebook M's general-purpose answering machine got shut down. So if even Facebook does not want to handle chatbot queries in any context, then maybe you shouldn't, too. At least for now.

Let's look into the more generalist bots. Microsoft's XiaoIce is impressive, but you need a Microsoft-caliber research team to build a bot with such a strong understanding of the world. As for Mitsuku, she has been under development since 2005, and while she is engaging, she is not as engaging as XiaoIce, which boasts that the average person who adds XiaoIce talks to her more than 60 times per month.

Find a valuable area for your bot

It is difficult to market your bot given the intense competition in the space. Every problem has multiple bots trying to solve it! Even only on Facebook, there are already 30K+ bots. The successful bots really solve a specific problem like scheduling meetings or generating reports.

Select the right platform for your needs

There are numerous Natural Language Platform vendors that provide APIs and each has strengths and weaknesses. Make your research before selecting the platform.

Work with experts and test, test, test

Not every conversational bot is worth working with outside experts. However, if this is a strategic initiative for your company, bringing in experience about the topic can be beneficial. And invest in testing to avoid embarrassing mistakes. The KPIs of the chatbot is an important part of testing and bot management. Close attention to user behavior can help identify unintended behavior of the chatbot and improve its functionality.

Conversational AI testing

Chatbot success is elusive and claims such as 10 times better ROI compared to email marketing makes sense only if the chatbot is implemented successfully. Combination of pre-launch tests (automated and manual tests) and post-launch A/B testing customized for chatbots can help companies build successful chatbots.

What are the tests to complete before launching a chatbot?

Good developers build automated tests for the expected input/output combinations for their code. Similarly, chatbot's natural understanding capabilities and typical responses need to be tested by the developer. These automated tests ensure that new versions of the chatbot does not introduce new errors.

The three types of tests below (general, domain specific and limit tests) need to be completed and ideally automated before releasing the chatbot. They test the key points of chatbots and would enable a company to pinpoint problems before launching its chatbot. After chatbot launch, they need to be automatically repeated to ensure that the new version does not break existing functionality.

General Testing

This includes question and answer testing for broad questions that even the simplest chatbot is expected to answer. For example, greeting and welcoming the user are tested.

If the chatbot fails the general test, then other steps of testing wouldn't make sense. Chatbots are expected to keep the conversation flowing. If they fail at the first stage then, the user will likely leave the conversation hurting key chatbot metrics such as conversation rate and bounce rate.

Domain Specific Testing

The second stage would be testing for the specific product or service group. The language and expressions related to the product will be used to drive the test and ensure that chatbot is able to answer domain specific queries. In case of an e-commerce retailer, that could be queries related to types of shoes for example. An e-commerce chatbot would need to understand that these all lead to the

same intent: “cage lady sandal shoe,” “strappy lady sandal shoe,” or “gladiator lady sandal shoe.”.

Since it is impossible to capture every specific type of question related to that specific domain, domain specific testing needs to be categorized to ensure that key categories are covered by automated tests.

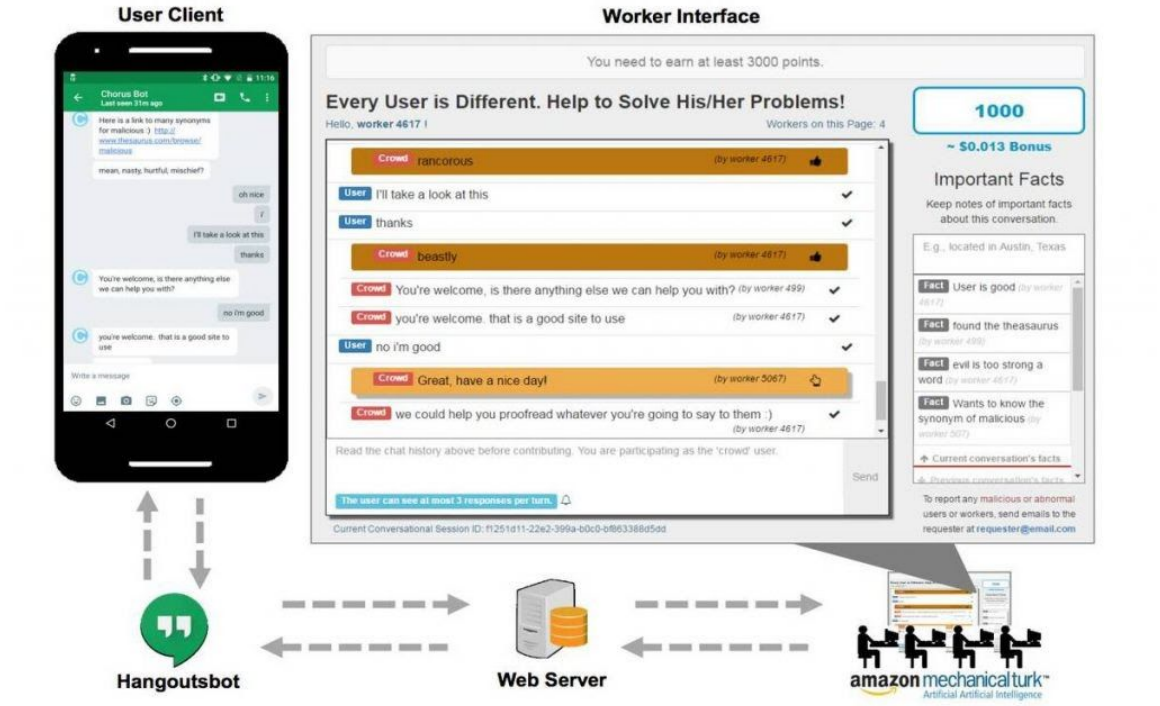
These context related questions will be the ones that drive the consumer to buy the product or the service. Once the greeting part of the conversation is over, the rest of the conversation will be about the service or the product. Therefore, after the initial contact and main conversations, chatbots need to ace this part or attain the maximal correct answer ratio, another key chatbot metric.

Limit Testing

The third stage would be testing the limits of our chatbot. For the first two steps, we assumed regular expressions and meaningful sentences. This last step will show us what happens when the user sends irrelevant info and how the chatbot would handle it. That way, it would be easier to see what happens when the chatbot fails.

Manual testing

Manual tests do not necessarily mean the team spending hours on testing. Amazon’s Mechanical Turk operates a marketplace for work that requires human intelligence. The Mechanical Turk web service enables companies to programmatically access this marketplace and a diverse, on-demand workforce. Developers can leverage this service to build human intelligence directly into their applications. This service can be used for further testing and reach for a higher confidence interval.



Source: Amazon Mechanical Turk

Conclusion

Our aim in this document was to demonstrate why you probably need a voice AI solution and how you should choose one.

The most important next step is to identify areas where conversational AI can add value. Feel free to reach out to us @ info@aimultiple.com if you need help in identifying how this technology can add value to your business.

Additional resources

- Natural Language Understanding in 2021: in-Depth Guide
- Natural Language Platforms: Top NLP APIs & Comparison
- Conversational User Interfaces in 2021: In-depth Guide
- Top chatbot testing frameworks & techniques in 2021

For more information, please contact info@aimultiple.com