DefinedCrowd

Training a Voice Assistant

From recognizing speech to understanding customer sentiment

by **Daan Baldewijns**
Director of Technical Program Management at DefinedCrowd

The Rise of Voice

Voice technology is transforming customer experiences. It's now more important than ever to be able to understand customers as they vocalize their wants, needs and preferences and expect their desired outcome in an instant.

Once a novelty, voice-enabled assistants are now an everyday presence in our lives – whether they reside in the operating system of a mobile phone, the dashboard of a vehicle, or in the intricate wiring of a smart speaker. Not only are they more commonplace, they are increasingly expected to work beyond the basics. Templated answers that miss the mark, or phrases such as, "I'm sorry, could you repeat that", simply don't cut it.

Today's consumers have less time and higher demands than ever before and in order to retain their interest, loyalty and share of wallet, businesses need to start thinking about voice as part of their strategy. The key ingredient? High quality training data to enable fluent conversations that exceed expectations.

> **"Businesses need to start thinking about voice as part of their strategy. The key ingredient? High quality training data."**

It's estimated that by the end of 2020 50% of all search will be conducted via voice and 75% of homes in the U.S. will have at least one smart speaker. This will mean significant changes that will see businesses not only optimizing for voice, but creating whole experiences and content designed primarily for voice interactions.

As the evolution of voice technology continues to accelerate, it will be able to offer value not just for lifestyle and entertainment, but also for healthcare, education and essential day-to-day needs and activities. In order to deliver that value, the quality and range of speech training data used to power these systems will be instrumental in ensuring every voice is heard.

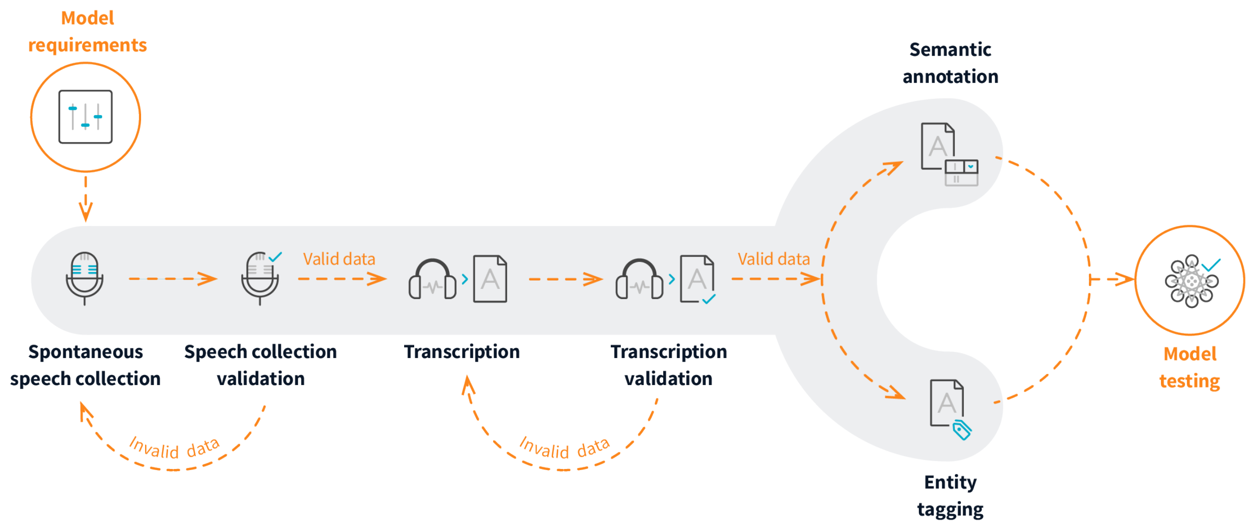
We created this white paper as an overview of how voice assistants are designed, trained and built, in order to enrich customer experiences and bring competitive edge to the businesses that create them.

Learning to Communicate

Building a system that can understand human speech – or an automatic speech recognition (ASR) model – is one thing. For it to be of value and service, it needs to do more than simply hear. It needs to listen, understand and respond. Just as each of these stages of communication requires more from a thinking, feeling human; so more complex training is needed for them to be replicated by a machine.

For a voice assistant to conduct fluent, near-human like conversations and enable smooth, helpful interactions with its users, it needs to be trained with data that is specific to its purpose. This data is sourced and structured through a combination of workflows that include speech collection, transcription, annotation and tagging, with various stages of validation along the way.

An overview of a speech collection and annotation process



Before delving into the details of those workflows and processes, let's walk through the basic requirements of hearing, understanding and responding as they relate to a voice assistant that is learning to communicate

Hearing

Before any real interaction can take place, a voice assistant must learn to hear. This requires the development of an acoustic model which can accurately pick up the physical emission of sounds. Of course, the world if full of many types of sounds. Not all of them are human and of those that are, not all are speech.

The acoustic model needs to detect those that represent language, such as vowel and consonant sounds or "phonemes," and be able to string those sounds into words. For an ASR model to have the highest chance of success, it should be trained with data tailored and tuned to the specific scenario it will be operating in, taking into account any environmental factors and the multitude of competing noises that could effectively impair its hearing ability.

Example 1: In-car infotainment

A voice-enabled car dashboard will need to be able to hear and respond to the driver through a lot of unforeseen background noise, from traffic sounds, rain or thunder to other passenger voices within close range. It will therefore need specific training to hear its user in as many environments as possible.

Example 2: Home smart speaker

When it comes to voice assistants at home, once again the model needs to discern its instructions over the noise of anything from the kitchen blender to the neighbor's lawn mower. Simulating these environments when collecting the speech data will help ensure a better performance.

Additionally, it needs to distinguish sounds such as coughing or vocal pauses ("um" or "ah") from those that actually form parts of words. For this, the acoustic model needs to be paired up with a language model so that it can accurately transform those sounds and words into meaningful sentences.

Understanding

Once the model can effectively hear customers' words and phrases, it needs to be able to understand the meaning behind them. This requires building a system that can recognize **domain** and **intent**.

Learning to Communicate

Simply put, domain is the context, subject matter or frame of reference relevant to the voice assistant. Examples of domains can be anything from entertainment, travel, or automotive, to banking, healthcare or food. Intent is somewhat self-explanatory: it's what the user *intends* to do, learn or achieve through his or her request, whether it's playing music, cancelling a credit card, asking for a prescription or finding the nearest gas station.

A voice assistant model must be fed with enough data pertaining to the domains it serves and enough examples of relevant customer intent, so that it can understand the requests, complaints or commands they are most likely to receive.

This requires the sourcing of training data enriched with semantic annotation, named entity tagging and other data structuring processes designed to "label" elements of data in order to derive meaning.

Responding

Having understood the commands of its user, a voice assistant then needs to respond appropriately. This may happen through the mapping of certain requests with pre-scripted answers (think IVR systems for call centers), or with actions such as playing music (think Spotify). If you literally want the model to talk back to the customer, that requires another phase of programming to give your device its own voice, as in the case of an Alexa or Cortana.

In the context of customer service, there may be times when it's best that a customer is passed on to a real human for support. This can happen automatically if a customer takes no action or selects the option to talk to an agent. However, an even more effective IVR assistant might also detect if a customer is getting angry or impatient and proactively escalate the issue and move the customer to the front of the queue for person-to-person resolution. For this, sentiment analysis data can be used to train a model to pick up customers' emotions from the way they communicate.

In the next sections we'll go through the key workflows and processes, both industry-wide and those we use at DefinedCrowd, for developing and evolving a voice assistant with high-quality training data.

Speech collection

The basics of speech collection may seem simple enough: ask some people to record some phrases, then use those recordings as the data to train the models behind voice assistants. But how does it all happen?

The right voices

In order to train a model to understand customers, you're going to need samples of speech from people who represent them. That means sourcing a crowd of humans who match the linguistic and demographic profile of your target customer group. Typical basic criteria include but are by no means limited to:

- Age
- Gender
- Native language/ dialect
- Country/ region of origin

Once a large enough group of the right set of speakers are sourced, they are tasked with specific instructions to create voice recordings which will then be used as the raw speech training data for the voice assistant model.

Different types of speech data

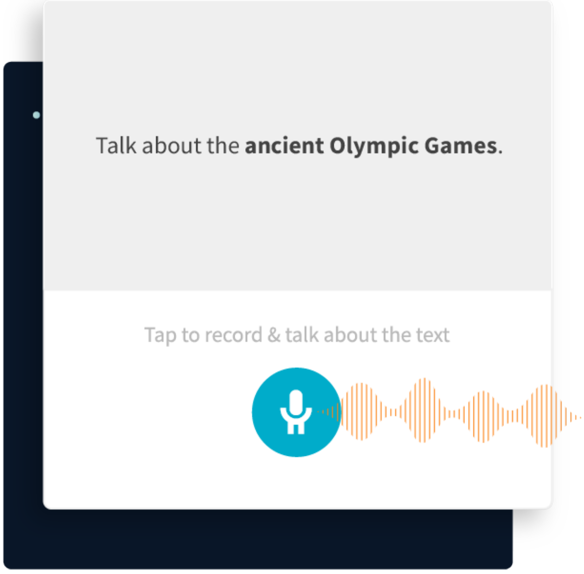
The recruited people are asked to record phrases related to specific scenarios pertinent to the domain of the project, for example banking inquiries or in-car entertainment. Depending on the desired specifications of the model being built, different types of speech data could be required, for example:

Scripted speech

For very specific uses cases, such as an IVR designed to deal with particular customer inquiries, having speakers read a pre-written script of phrases or questions can be an effective way to gather the training data needed. This might include things like, "What is my bank balance?" or "I want to increase my credit card limit". This is called scripted speech collection.

Spontaneous speech & variant collection

For scenarios where the end user of a voice assistant could make requests or commands in a number of different ways, it makes more sense to gather spontaneous speech data – where the speakers produce "utterances" in their own words. Here, participants are provided with instructions and prompts, such as, "Ask to listen to some music". This process captures more diversity and variants of speech. Due to its unpredictable variation, more data is usually required for spontaneous speech. It can be helpful to have a mix of both spontaneous and scripted speech to ensure as many scenarios as possible are covered.



Talk about the **ancient Olympic Games**.

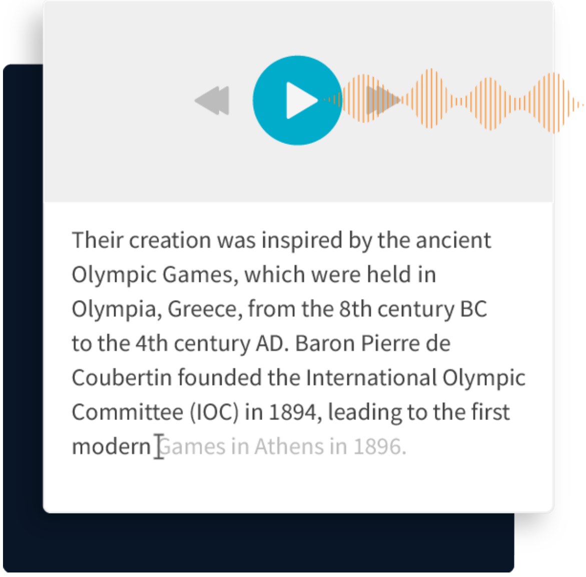
Dialogue speech

In the case of voice assistants for customer service, dialogue speech collection can be used to gather conversations between an agent and a customer. If a company has their own data from real recordings of calls, these can also be used to train the model, supplemented with additional data based on role play recordings. However, in the case of new product launches where real calls haven't yet taken place, recording pseudo scenarios and conversations in anticipation of customer queries means you're already set up to deal with them.

Transcription & validation

Once the speech data is collected as audio recordings, another group of people should then listen to and transcribe the data into text. These transcriptions will be used to train the ASR system and are therefore of the utmost importance. The more accurate the transcription, the more robust the ASR system will be. That's why, for a higher level of quality and accuracy, a new group should then validate those transcriptions, checking that they correctly capture everything – that means not only words spoken, but also pauses, coughs, hesitations and repetitions.

Data that doesn't pass the validation should be recollected and validated.



Their creation was inspired by the ancient Olympic Games, which were held in Olympia, Greece, from the 8th century BC to the 4th century AD. Baron Pierre de Coubertin founded the International Olympic Committee (IOC) in 1894, leading to the first modern Games in Athens in 1896.

Annotating and tagging

It’s one thing to identify words and phrases, but it’s the meaning behind them that counts, especially when it comes to delivering a great customer experience. Once a model can transfer speech into text, it then needs to know what to do with it. This is where natural language processing comes in.

Semantic annotation

After the speech data has been collected and transcribed into text, the validated transcriptions can then be annotated to reap greater value from the data by identifying the speakers’ intent.

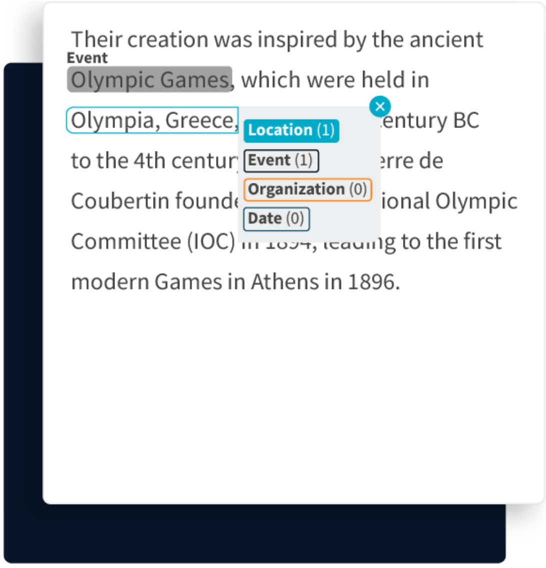
Depending on the use case of the voice assistant, categories would be defined based on the specific customer scenarios it will need to support. In the case of the in-car infotainment system, possible intents could include adjusting volume, playing a song, or even finding a restaurant. For a banking IVR system, they would cover things like making a transfer, updating an address or checking a statement.

Annotators go through the transcribed speech data and label different sentences or excerpts of text according to the pre-defined categories of meaning and intent.

Named entity tagging

This process involves a group of workers from the crowd or community once again going through the transcribed speech data and tagging anything that is a relevant entity for the given domain.

For example, “Lady Gaga” would be tagged as a “Singer”, and “jazz” as a “genre”, and so on, depending on the domain and the relevant categories.



Their creation was inspired by the ancient
Event
Olympic Games, which were held in
Olympia, Greece, entury BC
to the 4th centur erre de
Coubertin founde ional Olympic
Committee (IOC) ...g to the first
modern Games in Athens in 1896.

Once a model can transfer speech into text it then needs to know what to do with it. This is where natural language processing comes in.

As there can sometimes be ambiguity in a speaker's intent, there should be multiple annotators and quality processes in place to ensure the most objective assessment possible. DefinedCrowd uses sophisticated inter-annotator agreements to get better consistency and higher data quality.

Sentiment analysis

One of the biggest challenges of artificial intelligence is just that: it's artificial. Fully understanding and mimicking human thought and behavior will never be a flawless process, if only because as humans we are so diverse and unique. So, when it comes to emotional AI – training machines to understand and respond to feelings and moods, as well as deciphering the subtleties of opinions expressed with irony or sarcasm – it's a whole other level of intricacy in the development of an ASR model.

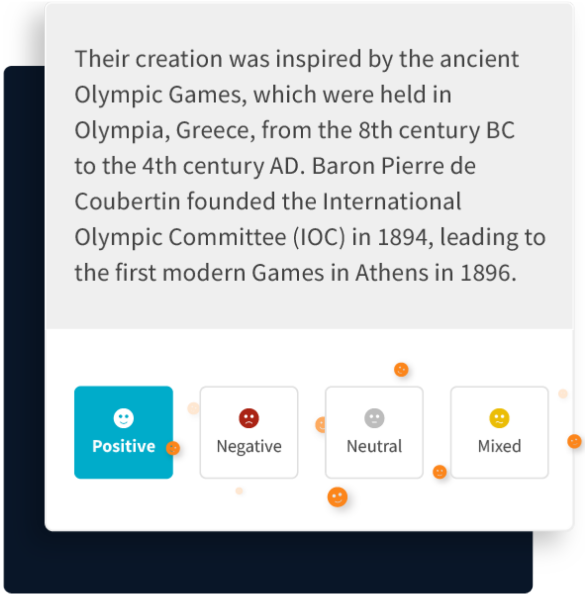
Detecting human emotion

Analyzing sentiment in speech transcriptions offers great potential for businesses that want to ensure a seamless experience at every stage of the customer journey – especially when things go wrong.

The last thing an unhappy customer needs is a robot misunderstanding them; and the last thing brands need is that customer taking to social media to put a blemish on their reputation. By detecting a customer's mood and sentiment, voice assistants can be trained to deliver the right message in the right tone, or take appropriate actions. So how are they trained to do this?

Once again, a group of people need to be tasked with annotating data, this time with a focus on emotion, mood and sentiment. That means picking up on when a person is becoming impatient, angry, confused or frustrated; or when they are expressing dissatisfaction with a product or service. Due the subtle nuances in speech, this can be a big challenge for a machine. Think of phrases like, “Well that's really made my day”. In isolation, this could sound like a delighted customer; but in the context of a lengthy call that ends in bad news, it could essentially mean, “this has been the worst experience ever”.

Understanding your Customer



Their creation was inspired by the ancient Olympic Games, which were held in Olympia, Greece, from the 8th century BC to the 4th century AD. Baron Pierre de Coubertin founded the International Olympic Committee (IOC) in 1894, leading to the first modern Games in Athens in 1896.

If a voice assistant can read between the virtual lines of this type of speech and pick up on nuances, its capabilities can then extend to escalating an unhappy customer's issue and getting them person-to-person help before the situation spirals. On the other hand, it could also be trained to tag a happy customer as a potential brand advocate, or a candidate to trial a new product or participate in a survey. If used in the right way, a voice assistant can do more than simply respond to commands, it can also help provide insights for marketing, product development and more.

Key Considerations for a Voice Assistant

If you're looking to build a voice assistant, it's essential to define the specific requirements and scenarios it will need to cater for, to get the right result. The following questions are key:

1. **Do you provide SLA's for quality and throughput in contracts?**

   The answer should be yes. That means they're serious about data quality. What's more, they should be able to provide a comprehensive overview of how they track quality and which quality measurements factor into their exit criteria. Otherwise their "guarantees" don't mean a whole lot.

2. **What languages do you support?**

   It's prudent to delve into a firm's language capabilities beyond the scope of your immediate needs, particularly if your company operates on a global scale. You want to avoid the headaches, and costs, of dealing with multiple vendors spread across geographical locales. Your data partner needs to be able to handle both your current and future needs.

3. **Do you provide both self-service and premium data capabilities?**

   Self-service platforms are great for quick-turnaround jobs you can handle setting up on your own. Premium, custom data collection and annotation may take a bit longer but will play a key role in new product development.

4. **Who are your current clients and partners?**

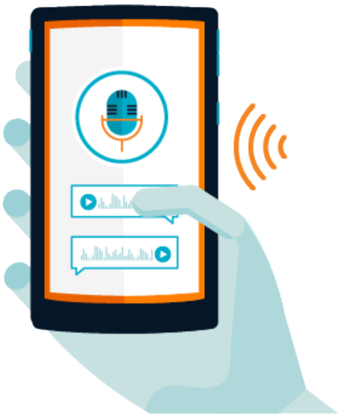
   If they're not already working with companies you recognize, you're taking a risk.

1) What environment will it need to operate in?

Depending on how and where customers are going to interact with your voice assistant, it will need to be trained to **hear through the noise** that is likely to interfere with its interactions. This could be anything from traffic and sirens to severe weather, machinery or even animal sounds, as well as the presence of cross-talk. Another key factor is how customers will use the assistant – whether via a desktop or mobile app, or with more specific device such as an Amazon Echo which involved them speaking from a distance. When there is a combination of both distance and spontaneous noise interference, the challenges to the voice assistant's ability to hear will be multiplied. All these considerations should shape the criteria for your training data needs.

2) Whose voice(s) should it hear?

An effective voice assistant needs to be able to communicate with the right end user. This means defining criteria such as age and gender in order to source the **right speakers** to create the speech training data. Language is of course critical, as well as country of origin. If your core market is in Australia, for example, your model will need to understand specific local phrases that may not be used by other native English speakers. In addition, a good voice assistant should be able to recognize and distinguish between different speakers, to serve the right user's needs at the right moments. This is particularly important for assistants in the home, where they may pick up a number of different voices.



3) What type of content will it need to understand?

Apart from defining your actual target language (Spanish, Japanese etc), this is about **domain** and **intent** and the specific lexicon related to them. If your domain is banking, the assistant will need to understand words and phrases expressing "intents" such as activating cards or getting a balance statement. For an entertainment system, the assistant will need to register commands like increasing volume etc. However, small details in expression can alter the meaning of a request. Picking up the term "credit card" but failing to understand the problem can be highly

Key Considerations for a Voice Assistant

frustrating. Imagine a customer whose card was just stolen, calling urgently to cancel it but receiving an upsell message in the process. Hardly a delightful experience.

4) What level of interaction will it offer?

Will the assistant only exist to solve a limited set of problems, or should it almost stand in for a human? While it can be tempting to create a basic model that ticks the on-trend AI box, if customers are going to spend time interacting with an assistant only to hear the response, “I’m sorry I don’t understand”, its good intentions could end up backfiring.

Whatever type of voice assistant you’re looking to create, it’s important to acknowledge some of the challenges and have clearly defined requirements in order to bring it to life. By thinking carefully about how it will serve your user needs, ensuring you get the right high-quality training data, and positioning it in a way that sets the right expectations; an effective voice assistant can help improve sales, call center efficiency and brand engagement, as well as delivering a whole new level of customer experience.

A Final Note

Adding value through voice

Voice technology is changing the way we all interact with the world. As consumers have less time and higher expectations than ever, businesses that want to compete for their loyalty and provide a superior customer experience can look to bring speech-enabled assistants into the mix. From helping resolve customer issues to recommending products, nurturing long-term relationships and creating brand ambassadors, building a voice assistant is one of the ways you can get closer to customers and take your conversations – however virtual – to the next level.

Chat to Us (Without the Bot)

DefinedCrowd was born out of a data scientist's frustration with first-wave data providers' inability to provide the high-quality structured datasets needed to push the technology forward.

Our first-of-its kind platform combines machine learning quality assurance procedures with cutting-edge human-in-the loop annotation practices, allowing us to be the only training data service provider in the field that includes SLA's for quality and throughput directly in our contracts.

We're proud to be supporting visionary companies, from Fortune 500 businesses to emerging tech leaders, who see the huge value and potential of AI looking to deliver results and revolutionize customer experiences.

If you want to understand more about how AI can play a part in your roadmap, and what type of data you need to get there, we'd love to talk.

Just tell us what you need:
sales@definedcrowd.com

How our platform works:

1. **Select your data source**

   Whether you have data to start with or not, you can create and configure a project with your own data parameters and rules. We can source the data for you or structure the data you provide – or a combination of both.

2. **Track progress**

   You'll be able to monitor progress of your projects while our platform combines human-in-the-loop annotation procedures with machine-learning powered quality assurance, to deliver precision results.

3. **Receive high-quality data for your specific requirements**

   You'll receive ready-to-use high-quality training data to fuel your models and reduce your time-to-market. Our services cover the full end-to-end process, from modeling, tuning, and expanding language capabilities, right through to deployment.

DefinedCrowd's first-of-its kind platform combines human intelligence and machine-learning backed quality assurance to deliver the quality-guaranteed, project specific data necessary to successful AI initiatives.

Want to know how a quality-focused training data partner will improve your products?

We'd love to talk:
sales@definedcrowd.com

DefinedCrowd