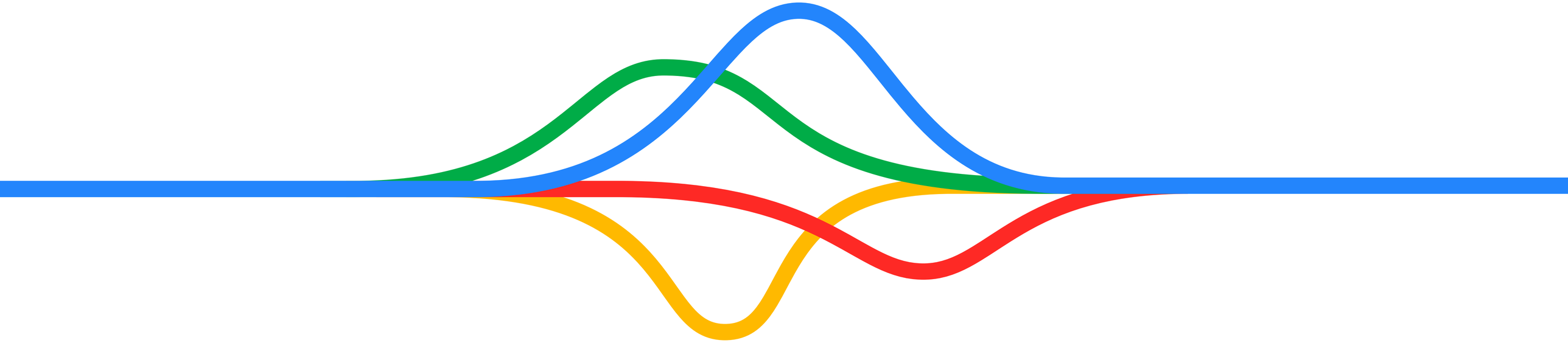
Google

Voice Playbook

Building for the **Next Billion Users**

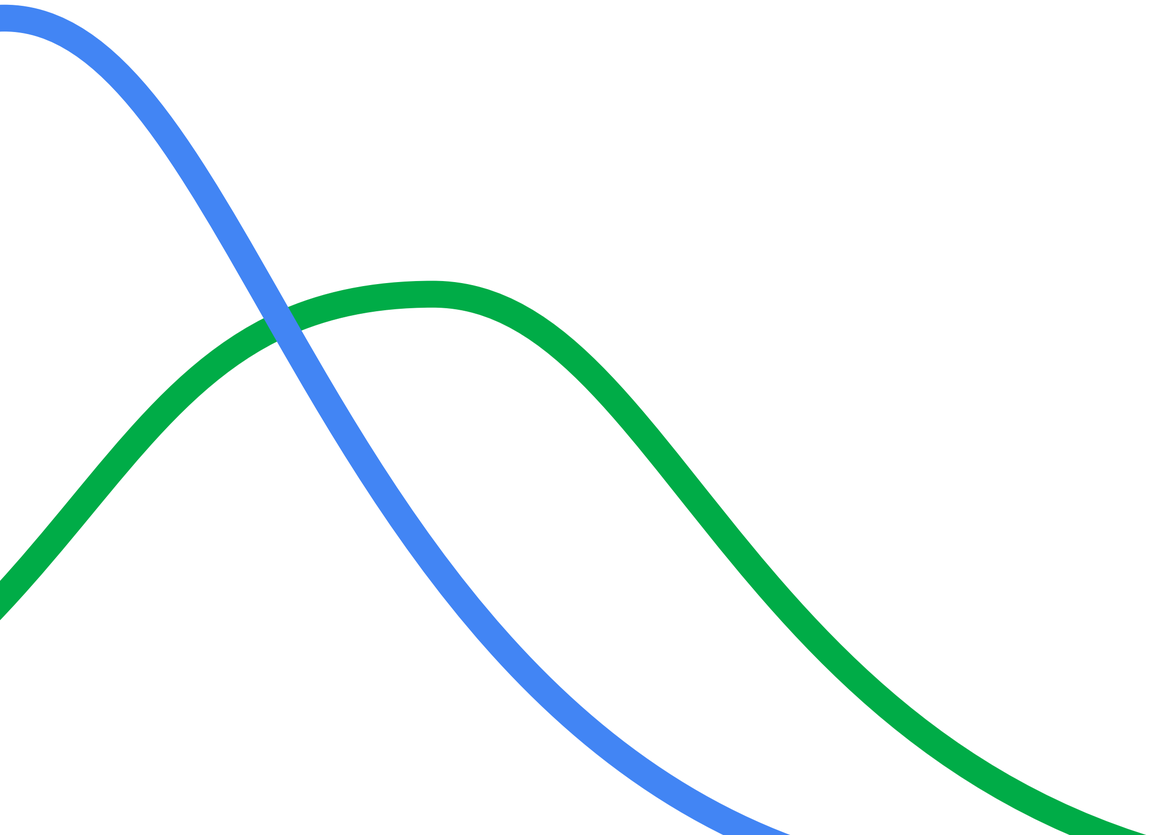


nextbillionusers.google

JUNE 2021

Google

Table of contents



INTRODUCTION

Voice, what's the big deal?

Voice has proven to be an increasingly essential interaction method for new internet users. Several studies have shown an uptake in the use of voice technology, and across Google products alone, we see a steady climb in voice usage traffic.

More than a million new internet users come online everyday, many of whom wouldn't be able to interact with technology at all without voice input. It may seem obvious that voice is a helpful interaction method for moments when you're unable to type, like when you're driving a car or are cooking in the kitchen. In reality, voice is not only helpful — it's critical— for many people's daily engagements with technology.

Voice has the incredible potential to reshape industry landscapes, but poses extremely complex and amorphous design problems. In this Playbook we have synthesized some of our key learnings from our own studies and interaction with new internet users, sharing a set of principles to help build more successful voice experiences for everyone.


Google
WHY VOICE?
Why are people using voice? It's empowering.

WHY ARE PEOPLE USING VOICE

Increased Self-sufficiency

Many **new internet users** struggle with lower literacy and depend on voice to interact with technology.

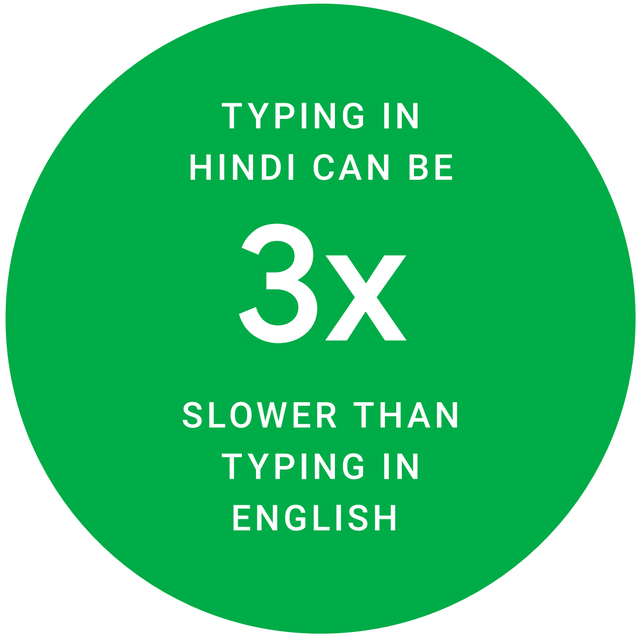
Voice empowers many users to do things themselves, without the assistance of others, increasing confidence while exploring the internet.

Simplified input

Typing in indic and scripted languages is a lot more complex and slower than in latin characters, giving voice an immediate advantage to users of all literacy types. However users at lower literacy levels get greater benefits from voice input by removing spelling complexities and making the process easier.
Understanding English is aspirational and many device language settings are often set to English by default. With voice input, new internet users can use their devices even if they can't read, write, or type in English.

Voice is especially beneficial for lower-end devices that use traditional T9 keyboards.

Sending voice messages is also super common, this makes mixing languages possible which is inherent in how many people speak and when typing, switching languages is cumbersome.



Understandable output

Output is an important part of voice functionality that is often overlooked. Many new internet users struggle with written content and the overwhelming majority of web content comes in the form of text. Many users wouldn't be able to use their device effectively if text-to-speech (TTS) hadn’t been read outloud to them.

Ability to multitask

The ability to multitask is a benefit of using voice technology for internet users globally. Many people we've spoken to have expressed that they are often juggling personal chores, household work, and multiple sources of income. Multitasking is common and voice functions in devices can help enable that.


Google
WHY VOICE?
Why aren’t people using voice? It can be frustrating.

Misinterpretation

Many new internet users blame themselves for a failed voice experience. Voice interpretation and speech recognition technology is not perfect yet, causing misinterpretations to happen to everyone, however new internet users often blame their sophistication and experience level instead. Almost every internet novice we have met expressed their frustration that it "didn't understand my accent." After a few poor experiences, they are way more likely to abandon using voice input.

Furthermore, there are multiple voice interaction types that are used today (e.g. dictation, commands, conversational, recording) and it's not very clear why and how they differ. This makes transitioning between them complicated and confusing.

For example, many, if not most, mobile device users hold their devices too close to their mouth when using voice, assuming it will help with word detection. They tend to hold their phones awkwardly and end up not looking at the screen, making it tough to see if the transcription is correct during text input. When the input is wrong, users get confused about how to correct an error and are constantly trying to fix misinterpretations with their voice by saying "pause" or "change 'water' to 'weather'". When this fails, those who can, resort to typing but it doesn't solve the confusion and implies to them that voice isn't as supportive as it should be.

WHY AREN'T PEOPLE USING VOICE

Self-perception and Privacy

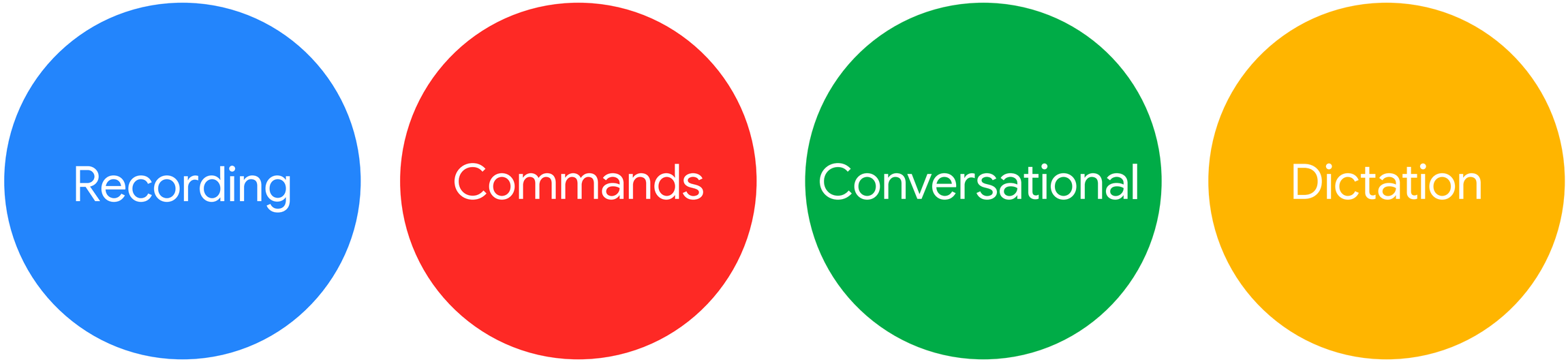
There is a widely accepted notion that voice tools help address illiteracy. Many users expressed their fear of being seen using voice because it could make them seem uneducated, or that their friends would make fun of them.

Another inhibitor is that people find themselves in situations where they are constantly surrounded by other people and are worried about others overhearing them and privacy.


Google
The current
state of voice.

Voice Interaction Types

There are four main voice interaction types that are widely used today. They are all built on drastically different voice interaction models and they may or may not be prompted by the same microphone icon, giving rise to a lot of confusion around the mental models of voice.



THE CURRENT STATE OF VOICE

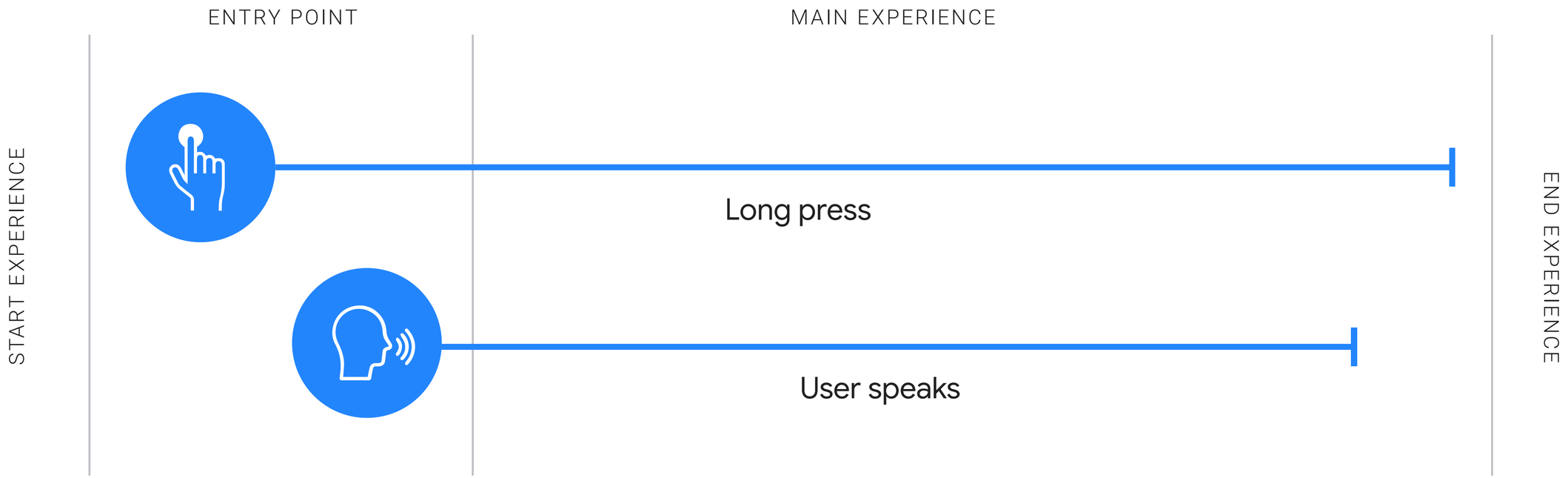
Recording

(tap and hold, walkie-talkie style)

NOTE This is the most widely understood interaction in NBU markets

TYPICAL USE CASE Messaging apps (WhatsApp, Messenger, Instagram DM)

ACTION Share a recording of the users voice

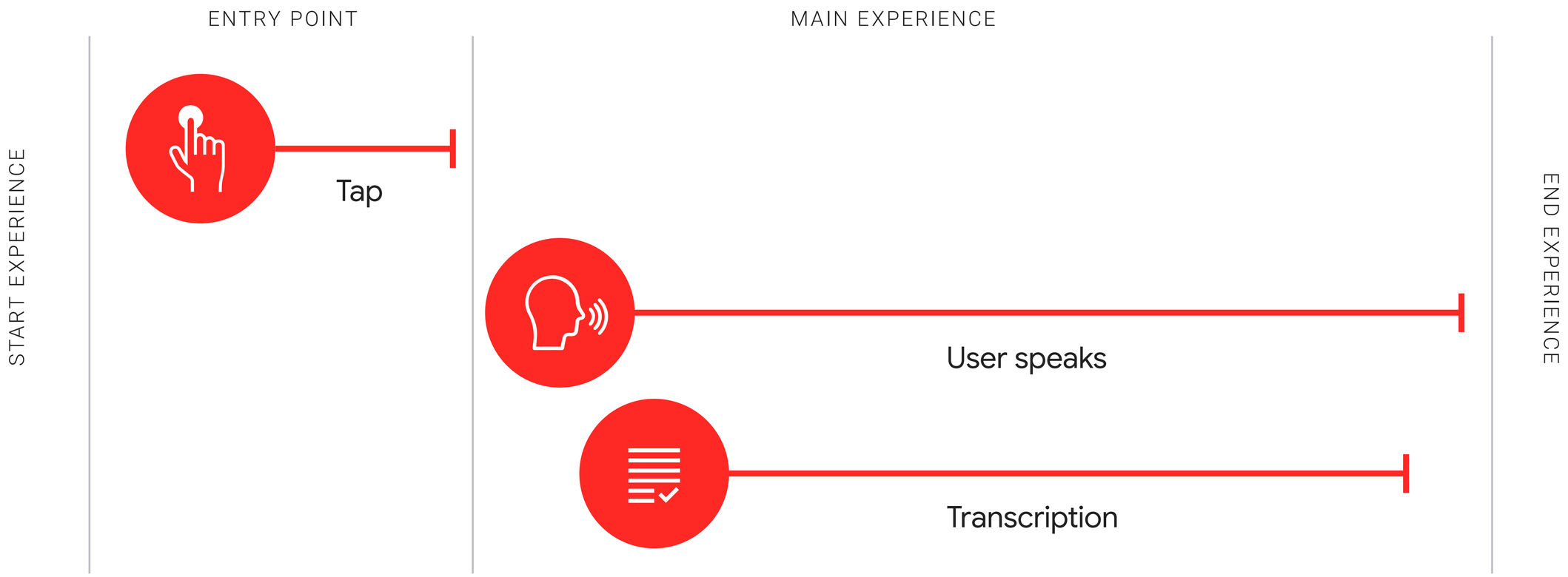


THE CURRENT STATE OF VOICE

Commands

TYPICAL USE CASE Search experiences (YouTube, Google Search)

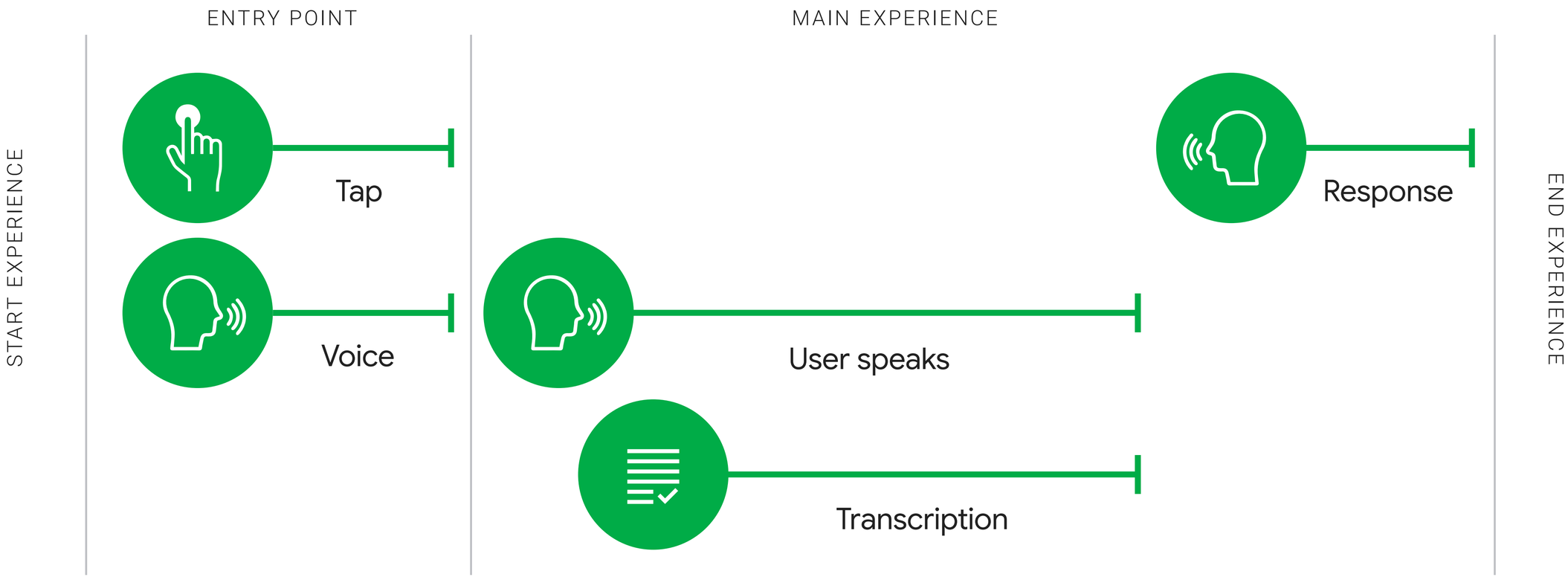
ACTION Users says word or phrase as query and receive result



Conversational

TYPICAL USE CASE Assistants (Google Assistant, Siri, Bixby)

ACTION Conversationally interact with devices. Users say commands and listen for a response.

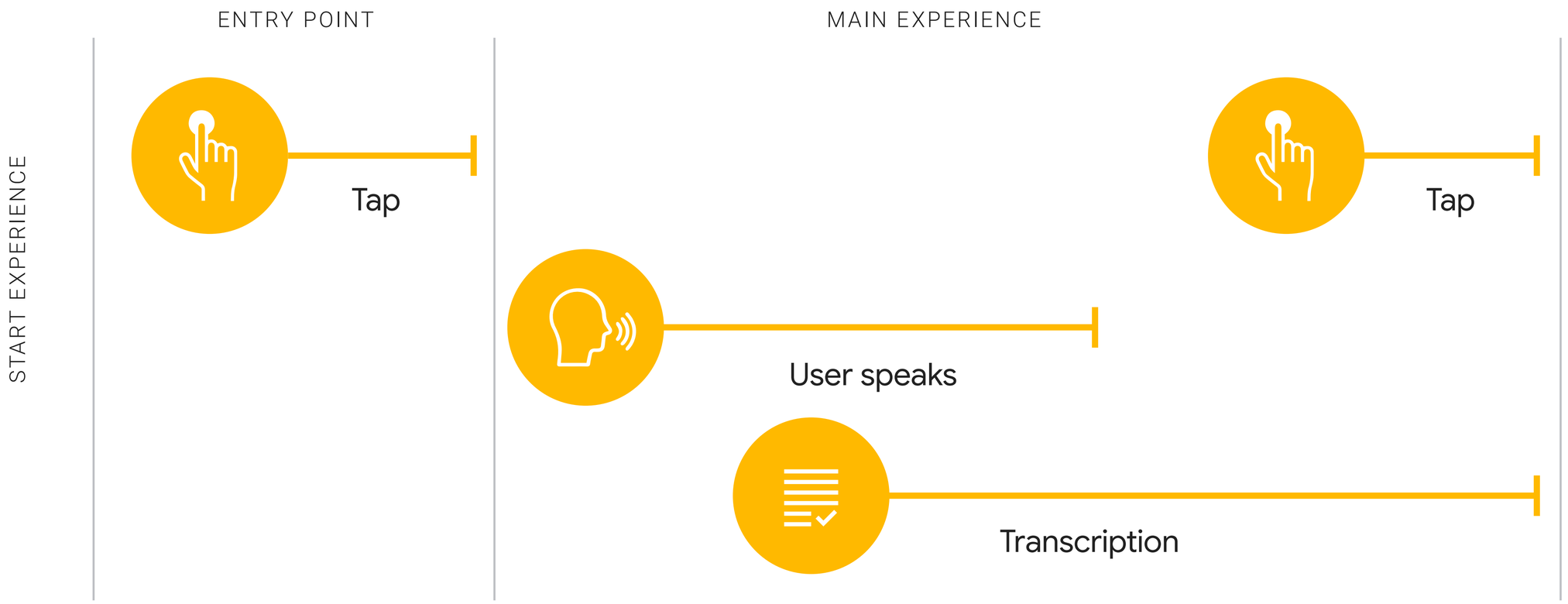


THE CURRENT STATE OF VOICE

Dictation

TYPICAL USE CASE Word editors, Text input, Keyboards

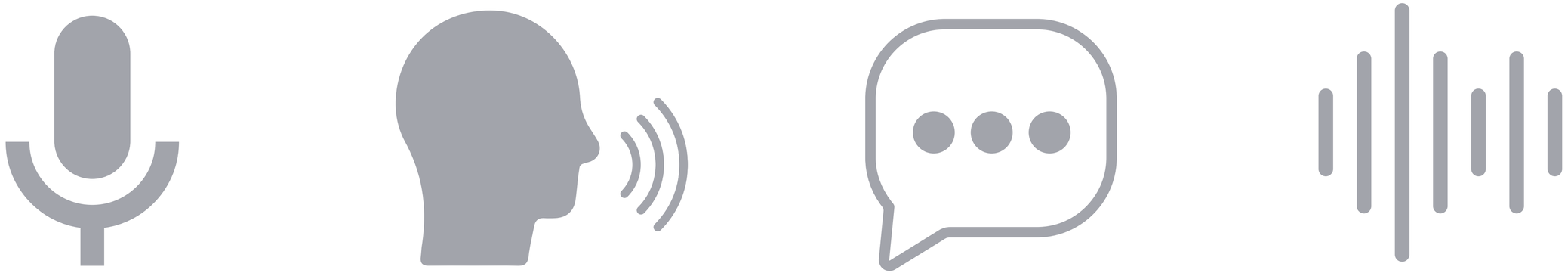
ACTION Users says word or phrase and it gets transcribed as text



THE CURRENT STATE OF VOICE

Voice Entry points

Not all voice experiences are represented by the same icon across different apps and devices. This creates confusion for users around expectations and even privacy. Not only does it muddy the mental model but creates visual problems too.



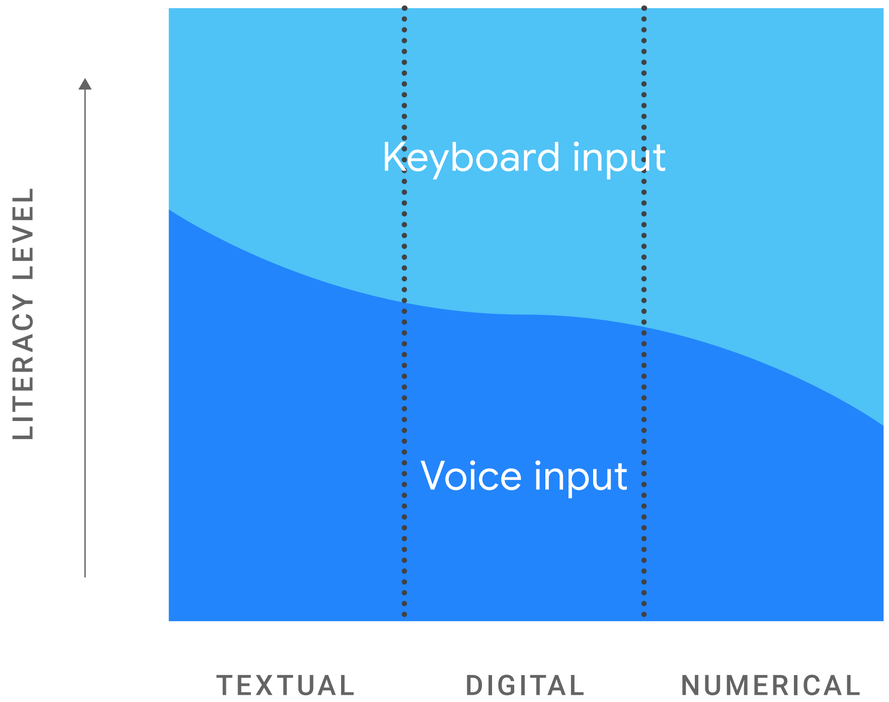
Keyboard vs Voice Input

Voice is helpful to many users across all literacy levels. However, voice is especially helpful to those who have lower literacy levels and we have seen those who have lower textual literacy resort to voice.

As users get more textually literate they are more likely to use the keyboard although this also varies across locations.

Many of these users switch to the keyboard when voice fails.

Keyboard vs. voice effectiveness across literacy level



Google

THE PRINCIPLES

7 Principles to tackle common challenges.

Google

7 Principles to tackle common challenges

We have identified 7 Principles that can serve as a guide on how to think about voice and how to ensure it will be an empowering tool for all users.

#1

Build for the user's reality

#2

Think beyond input

#3

Serendipitously educate users

#4

Capitalize on the familiar

#5

Make voice discoverable

#6

Design for errors

#7

Support multilingualism

Google

PRINCIPLE #1

Build for the user's reality

Consider the many situations and contexts in which voice interactions become handy for the user. Voice interactions are useful beyond simple hands-free or multi-tasking activities. The ability to bypass a keyboard or communicate via voice will be especially helpful to new internet users.

Google

Build for the user's reality

THE CHALLENGE | THE OPPORTUNITY

Unrecognizable Iconography

We ran a study in India and Indonesia on the voice icon, often represented by a microphone, and its comprehension. Throughout our research, we have seen many users failing to understand what the icon was.

When we asked users what they thought the microphone icon meant, many users were confused -- we heard “I think it’s for singing,” or “it’s for recording.” The 50’s style design is antiquated and unrecognizable to many new internet users.

However, when we asked about the more literal speaking icon, users very quickly understood what action they would be taking once clicking the icon. One said, “Oh, I say something to my phone. I like it because it shows me what to do.” The speaking icon also works to educate the user where the microphone does not.

Google

Build for the user's reality

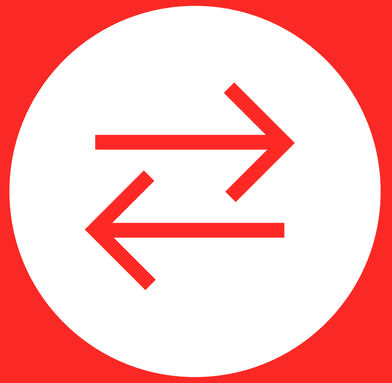
THE CHALLENGE | THE OPPORTUNITY

Use icons to teach the user about the expected action

Icons should help set expectations about what is going to happen after it is tapped/clicked. If possible use the icon to teach the user about the incoming interaction. Many users struggle with abstraction, so consider being more literal about the iconography.

Icons need to do a better job of telling users what they mean and what the expected action is.

Google



PRINCIPLE #2

Think beyond input

Offer flexible forms of voice interaction whether for input, output, or a combination of both. There are many useful forms of voice interactions that range from dictation, to voice recording and simple playback.

Google

Think beyond input

THE CHALLENGE | THE OPPORTUNITY

Timing of speaking or pausing

Sometimes, responses (text to speech) are presented in long sentence forms and can add cognitive load, making it difficult for users to follow especially if the user is not used to voice yet. That also makes it difficult for them to respond or answer to continue the “conversation.”

Google

Think beyond input

THE CHALLENGE | THE OPPORTUNITY

Make output length manageable and digestible

Artificial voice outputs should be equivalent to one human breath's worth of words. Knowing when to speak and when to listen is key in communication and ease of flow. This is especially important for low literate users who can't read transcription (i.e control to signal pause, listening and playback cues).

For voice outputs, if text-to-speech is being read out to the user, limit word count. Add pauses to break a lengthy sentence into smaller chunks. To test, have someone read the dialogue out loud to get a better idea of what the user will hear. This is especially important for lower literacy and multilingual users as it may take them slightly longer to parse incoming information.

Google

Think beyond input

THE CHALLENGE | THE OPPORTUNITY

Make output length manageable and digestible

For voice inputs, proper integration of microinteractions add clarity to overall user experience. Think about the time users may be taking for a quick pause to breathe and fully think through a query. Ensure that the microphone is open long enough for users to take a breath, and don't close the mic too fast. If it makes sense in the context of your product, provide the users the ability to "stop" or "pause" voice recording or voice input.

Google

PRINCIPLE #3

Serendipitously educate users

Many new smartphone users learn one or two voice actions and then their usage plateaus. This is caused by a fear of trying new things. Think of embedding the experience with support to help serendipitously educate the user and provide a continued onboarding experience.

A new internet user won't remain at beginner level forever. As they progress through their learning experience, getting more acquainted with their device, we need to continuously teach them to get more comfortable.

We see usage plateau when users aren't learning new inputs. Failure becomes a continuous occurrence which affects a user's confidence in trying new things.

Google

Serendipitously educate users

THE CHALLENGE | THE OPPORTUNITY

Input failure causes users to speak unnaturally

It can be difficult to formulate a search query, especially for users who are searching on their own for the first time. Many of them enunciate and speak louder in hopes of being interpreted correctly. Many of them also end up changing the way they speak and say phrases they may not naturally say. We observed that this is a learned interaction from input failure rather than the initial behaviour at first attempt. Once they fail in getting the specific result they're looking for, they change their voice input, thinking they did something incorrect, and then they end up having difficulty thinking of the right words to say or forming the best structure for their query. This causes even more failure in output.

Google

Serendipitously educate users

THE CHALLENGE | THE OPPORTUNITY

Educate, support and reinforce

We need to build a voice experience that educates users that they can speak naturally. Better support for these broken phrases needs to be in place while also teaching or guiding them on what is available to say. Suggestions can give users a starting point. Helping them discover available voice commands could also help provide a better learning curve.

At the same time, we need to make sure that the education doesn't alienate more experienced users.

Google

PRINCIPLE #4

Capitalize on the familiar

New internet users work hard to learn an experience and gradually tune in to a set of behavioral patterns. Once there is a developed understanding of an experience, users expect it to be consistent. When things unexpectedly change, they blame themselves or never come back to try again.

We learned that building upon interaction models that are already familiar to new internet users, and, when needed, taking care of easing the transition to new patterns, can be very helpful in making the experience better for new internet users. Ultimately, presenting familiar visual/voice/ touch interactions within apps will help new internet users to come on board easily.

Google

Capitalize on the familiar

THE CHALLENGE | THE OPPORTUNITY

Absence of standards hurt adoption and confidence

Currently, there are various ways by which apps present voice entry points. There is the ‘walkie-talkie’ or recording interaction used in some messaging apps, which is typically the first touch point for new users and thus widely understood in next billion users countries. When new internet users try other apps for the first time and the same actions (e.g., long press on microphone) don’t initiate the same experience (e.g., no voice recording), they are left confused. To a certain extent, they can also get scared thinking they broke something, or that something else happened that shouldn’t have happened, (e.g.,sending the recording to someone else).

Google

Capitalize on the familiar

THE CHALLENGE | THE OPPORTUNITY

Standardize iconography for easier adoption, or educate to facilitate switches across apps

Transitioning between voice experiences between different apps is confusing to users due to different affordances and interaction types, as stated earlier. We all need to do a better job in simplifying iconographies and interactions, using what is most familiar to clearly communicate the intended action. Similar to building confidence in users, providing support and education within the app to guide users on how to properly tap, press, or use icons is also a good way to familiarize them with different interactions.

Google

PRINCIPLE #5

Make voice discoverable

Voice interactions should be easily discoverable. If users don’t understand or can’t find the entry point they won’t be able to know where to begin. Aside from that, presenting an equal hierarchy between typing and voice can also prompt users to find more value in using voice.

Google

Make voice discoverable

THE CHALLENGE | THE OPPORTUNITY

The keyboard is often presented as the main interaction method and typically takes up 50% of the screen while the voice entry point takes up less than 1%. Inherently, this makes voice feel like the secondary option. Voice often is buried in the experience and not as accessible as typing.

Google

Make voice discoverable

THE CHALLENGE | THE OPPORTUNITY

Surface voice interactions immediately and prominently. Do not hide them behind multiple taps

Ensuring users have an easy way to initiate the voice experience is key. Bring out the voice interactions from behind multiple taps and give them a prominent icon placement that will not make it seem secondary to the keyboard.

Google

PRINCIPLE #6

Design for errors

We've all seen it in different scenarios — latency, misrecognition, and fail cases can happen with any interaction. Errors are especially likely in populations that have low connectivity and have linguistic complexities. In a lot of cases, users sometimes don't realize they have made a mistake until they fail to experience the expected, or for voice queries, until they see that they did not get the correct result. By then, the user is too confused to know how to correct the error or where to start again. However, the design and build for voice interactions leaves very little room for correcting the errors.

Google

Design for errors

THE CHALLENGE | THE OPPORTUNITY

Method of modification is not intuitive to voice

Once misinterpreted, users are eager to correct an error but often struggle to pinpoint exactly what went wrong. Even if they do realize where the mistake is, they have trouble finding out how to resolve it.

Correction is especially hard for low literate users. Transcription is currently the only indicator of an error which is not helpful to someone who can't read. Similarly, correcting an error can often be done only by modifying their phrase using the keyboard.

We've seen many users try correcting a mistake by using natural language saying "wait", "start over" "change ___ to ___". When that doesn't work they restate the phrase and hope they are "heard better". This leads to frustrated users who become more skeptical of using voice.

Google

Design for errors

THE CHALLENGE | THE OPPORTUNITY

Correction using voice

Our research has shown that users' first instinct when they encounter an error is to say "pause" or "stop" and try to tap the screen. They expect that if they input with voice they should also have the ability to control the experience, including correcting errors, using voice commands.

Potentially, the simplest way to allow users to correct an error is by showing a fully updated phrase that we assume the user intended to say. Using YouTube Correction Chips as an example, we found that the Query correction screen was well understood and perceived positively by all users(e.g., they know there was a problem encountered and they were prompted to make queries again). YouTube users paused to listen to the 'read aloud' on this screen. A plausible reason could be that it involved a 'refreshed state' and they were given the chance to check the correctness of the output and address the error.

Google

PRINCIPLE #7

Support multilingualism

Not all countries around the world are unified in a single language. India is a great example of this. People naturally merge multiple languages in a single sentence. Regardless of the global device setting, your users will potentially be using more than one language or a different language than initially set.

Google

Support multilingualism

THE CHALLENGE | THE OPPORTUNITY

Multilingual input and Language switching

A large challenge with voice is that the language of choice is pulled from default settings which many users set to English. However, many new internet users come from countries where they speak more than one language. During voice input, language switching makes it hard to understand resulting in misinterpretations in the output. However, we have seen that when users speak in their native language they are more likely to have a successful interaction, which helps in building their confidence in using voice.

Google

Support multilingualism

THE CHALLENGE | THE OPPORTUNITY

Provide multilingual or language switching support

The allowance of multilingual input could be really beneficial to users. Many users expressed that mixing languages is inherent in how they speak and the expectation is voice will be more supportive of that than typing.

Conclusion

There are another billion people set to come online in the next few years and it's our job to create accessible experiences that work for all. Voice has proven itself as a crucial interaction method for many of these new users, helping to democratize technology, regardless of the level of linguistic or technical literacy.

There are many users that currently depend on voice to interact with the world, however we are far from a perfect voice-driven future. In the meantime, we need to do our best to proactively solve for errors and educate users, as means to improve the experience.

How might we make voice feel like a **super power** rather than a crutch? How might we ensure everyone can feel heard by the technology they use?