SCTE CABLE-TEC EXPO 22
SEPTEMBER 19-22 • PHILADELPHIA, PA

2022 Fall Technical Forum
SCTE • CABLELABS • NCTA

Voice Control of Set-Top Box for Customers with Non-Standard Speech

A Technical Paper prepared for SCTE by

Adina Halter
Sr. Principal Software Architect
Comcast
1701 John F Kennedy Blvd, Philadelphia, PA 19085
267-658-0261
adina_halter@cable.comcast.com

Sara Smolley
Co-Founder, Head of Partnerships
Voiceitt
700 Canal Street, Stamford, CT 06902
716-348-8229
sara@voiceitt.com

Table of Contents

List of Figures

1. Introduction

Voice control of set-top boxes is becoming the norm. But voice technology needs to be able to understand non-standard speakers as well. Non-standard speech can be a factor for people affected by deafness, disabilities, medical disorders, or even foreign language speakers.

Comcast and Voiceitt, an Israel-based voice technology startup, have collaborated to explore a solution using a mobile application paired to a set-top box. In this paper, we describe how we are applying machine learning and artificial intelligence to create unique voice command models for individuals with speech disabilities to access the set-top box. In this paper, we will cover:

- Non-standard speech and how this translates to customer set-top-box control
- Solution method and architecture
- Description of our ongoing customer trial.

Offering accessible voice control for non-standard speakers can open the opportunity for all customers to experience the joy and convenience of a voice-enabled home entertainment system.

2. Non-standard Speech and the Set-top Box

This solution can be adapted to other devices besides set-top boxes (mobile, TV, computer, streaming adapter) using similar approaches and methods described throughout this paper.

2.1. What is non-standard speech

Non-standard speech is speech that is not readily understood by others or by standard speech recognition. This could be because the speaker has an accent, has deaf speech intelligibility issues, uses speech synthesis, or has a physical or neurological disability such as dysarthria. Other types of non-standard speech disabilities such as Wernicke's Aphasia create a disconnect between thought and utterance. The technological solution described in this paper will work if there is a regular, repeatable connection between the thought and the utterance.

2.2. Voice control and the set-top box

Voice-driven technologies are proliferating rapidly. Growing adoption of smart speakers and smart assistants is likely to make speech recognition a primary means to interact with the technological world around us, including home entertainment.

In 2015, the Emmy-award winning Xfinity Voice Remote Control introduced the ability to control a set-top box with one's voice. Using machine learning, Comcast's Natural Language Processing platform ensures that the remote delivers precise results. "The platform leverages machine learning to understand what customers mean when they say certain words or phrases and deliver highly relevant results."

The speech recognition engine requires understandable speech to perform speech-to-text (STT) conversion. Thus, people with non-standard speech cannot access mainstream SST voice technologies.

"Switch the channel to HBO"

For many of us, this simple, familiar voice command spoken aloud in our voice remote control is a convenient way to instruct our home entertainment system. For people with speech and motor disabilities, being able to use their voices would give them an opportunity to take ownership of their TV, increase their independence, and decrease their dependence on people around them to perform tasks such as changing the channel, recording a show, or browsing content.

In her report "Xfinity Adaptive Remote for Accessibility Audiences" Theresa Murzyn Ph.D., Lead UX Researcher at Comcast states “Not being understood takes a mental and emotional toll on users with motor/sensory challenges. The Adaptive Remote must enable users to feel understood regardless of their input method.”

Customers’ physical challenges, multi-sensory personas, and video mindsets inform what they expect from set-top box control technology, namely:

- Independence
- Fewer steps
- Speedier navigation through the TV/cable interface
- Less mental and physical effort
- Being valued

2.3. Current solutions for non-standard speech

“It doesn’t understand me. I don’t know why.” (giggle)
— Comcast customer with amyotrophic lateral sclerosis (ALS)

Many of the people who can benefit from “voice first” technologies cannot access those technologies because they do not have the standard speech patterns that are recognizable to commercial automatic speech recognition (ASR) algorithms.

So what options are currently available to them to navigate this "voice-first" world?

“I have the capabilities of doing streaming if someone else is here pushing the buttons for me. But I can’t do it myself. And so, I never do it.”
— Comcast customer with spinal injury

Non-standard speakers may rely on friends, family, or caregivers for basic tasks, including controlling their devices for everyday tasks. Voice control for their set-top boxes can provide independence in these everyday routines.

"I use my iPad for a lot of YouTube. And I used to use a [sic] voice activation to get to it, but now it doesn't understand me any longer. So, I've kinda' lost the use of it."
— Comcast customer with ALS

Technologies such as the Xfinity Adaptive Remote are beginning to give touch and text alternatives to those who cannot speak into a physical remote control. Customers can pair the Adaptive Remote with assistive technologies (ATs) such as eye control, mouth sticks, gross-motor options for swiping and large-target tapping. This pairing gives these customers an opportunity to trigger set-top box actions or submit a text string version of the "voice" command they are interested in. Submitting this text string bypasses the ASR algorithm.

And yet, this technology often requires many steps to complete a simple task. This makes content foraging slow, tiring, and often exasperating. As one of our customers with ALS remarked, "Every click is time."

"They are already managing many challenges. let's not add more to them."
— Theresa Murzyn, PhD.

2.4. Our approach

In this project, Voiceitt's non-standard speech recognition was integrated with the Xfinity Adaptive Remote's video code (vcode) and string input capability, using Comcast's Companion App architecture as the technology bridge. This enables customers with speech disabilities to access and control their set-top boxes (and wider home and entertainment platforms) by voice.

The solution presented here was the innovation of two companies, Comcast and Voiceitt. Please note that while we will often refer to our different technologies by company or product name in this paper, this is done to keep our two companies' individual contributions and solutions clearly differentiated. Similar solutions can be developed by your organization's product team as well.

3. ASR Technology for Non-Standard Speech

Voiceitt's ASR technology is designed to recognize the speech of people with speech disabilities. The technology includes both discrete and continuous ASR for non-standard speech. Discrete ASR offers the ability to recognize a predefined list of phrases which the user with speech disabilities can customize. For example, if the speaker trains the software with the vocal pattern "uhwuh o uhah" and its meaning is, "I want to go outside," the software learns to recognize this pattern and associate it with its meaning, which it can then produce through digital speech.

Continuous ASR, now in Beta, extends this functionality to recognize the user's speech more flexibly. With the continuous ASR, there are no longer constraints to use predefined phrases from a phrase bank, thus allowing the user to speak more spontaneously and freely.

Both the continuous and discrete ASR technologies are customized solutions tailored to the individual user. As such, they rely upon enrollment data (training data): samples of the user's speech. This enrollment data is used to adapt the acoustic model to provide a more accurate representation of the individual's speech.

Further, hands-free activation is supported using the Voiceitt wake word technology, extending further accessibility for users with disabilities.

In the case of the discrete ASR solution, this customized model is downloaded onto the edge device as illustrated below.

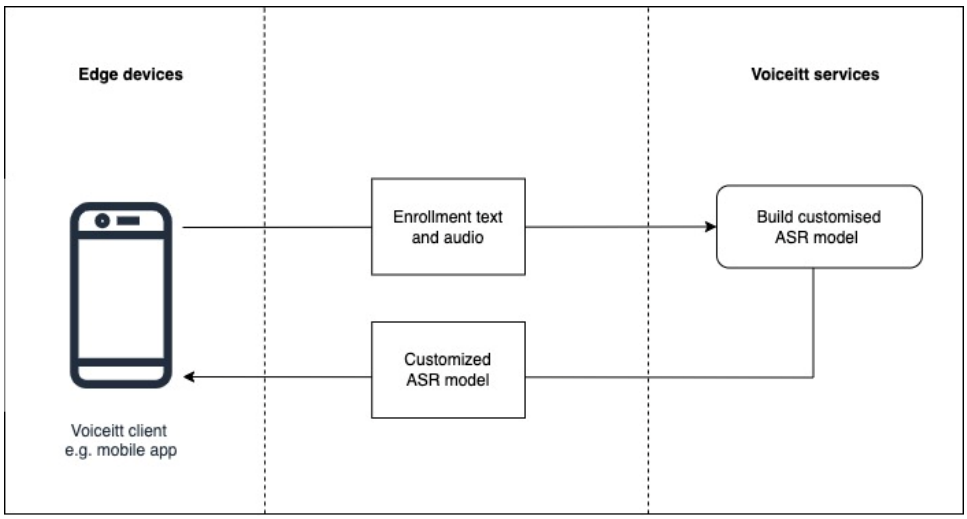

Figure 1 – customized model downloaded onto the edge device

Once the enrollment phase is complete, the user may use the edge device to recognize his/her speech. This recognition takes place on the device in the case of the discrete ASR solution. Importantly, a feedback loop is implemented which continuously improves the accuracy of the solution. This feedback loop is illustrated below.

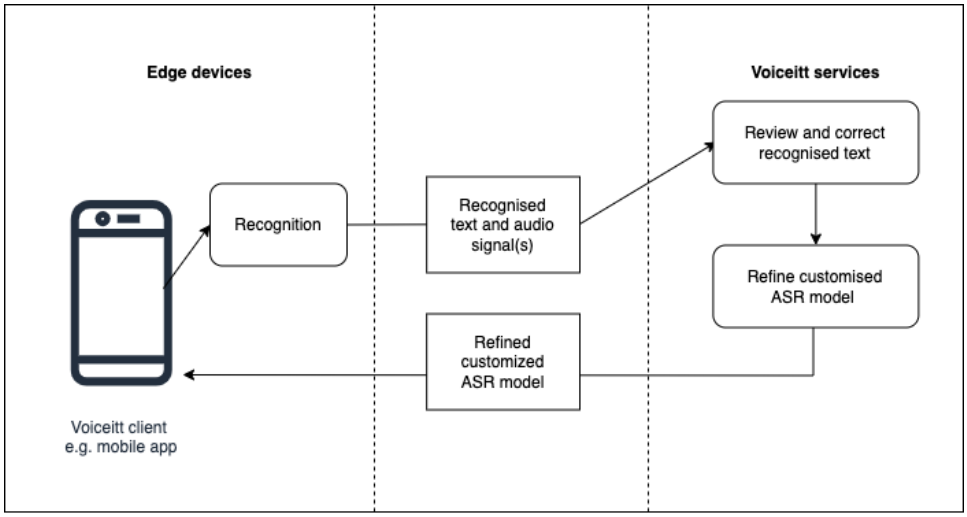

Figure 2 – feedback loop

Recognized text and its associated audio signals are sent to the backend database. The text and signals are reviewed and potentially corrected by a team of skilled annotators to provide additional training material to the model customization procedure. A further customized model is then delivered to the edge device for future recognition. This process iterates while the user engages with the technology, creating a virtuous machine learning cycle which delivers improved ASR accuracy.

A very similar workflow of enrollment, recognition and feedback is deployed in the case of continuous ASR. The primary difference is that continuous ASR uses a combination of an acoustic model and a language model to add the ability to recognize free speech using words and phrases that were not pre-trained, as well as phrases that are not in the pre-defined order.

This method could also be integrated with speech-based technologies such as those found on voice-controlled interfaces such as set-top boxes and smart home devices, effectively enabling users to access such technologies.

4. Adaptive Remote Technology

The Xfinity Adaptive Remote (https://remote.xfinity.com/) is a web application written in NodeJS which allows users to control their set-top boxes with various ATs such as the Tobii Eye Gaze solution.

The original project motivation was to provide remote tuning capability for our customers with ALS (also known as Lou Gehrig's disease) using an eye-tracking device such as the Tobii Eye Gaze. Later, features such as support for voice commands and voice-as-text commands were added.

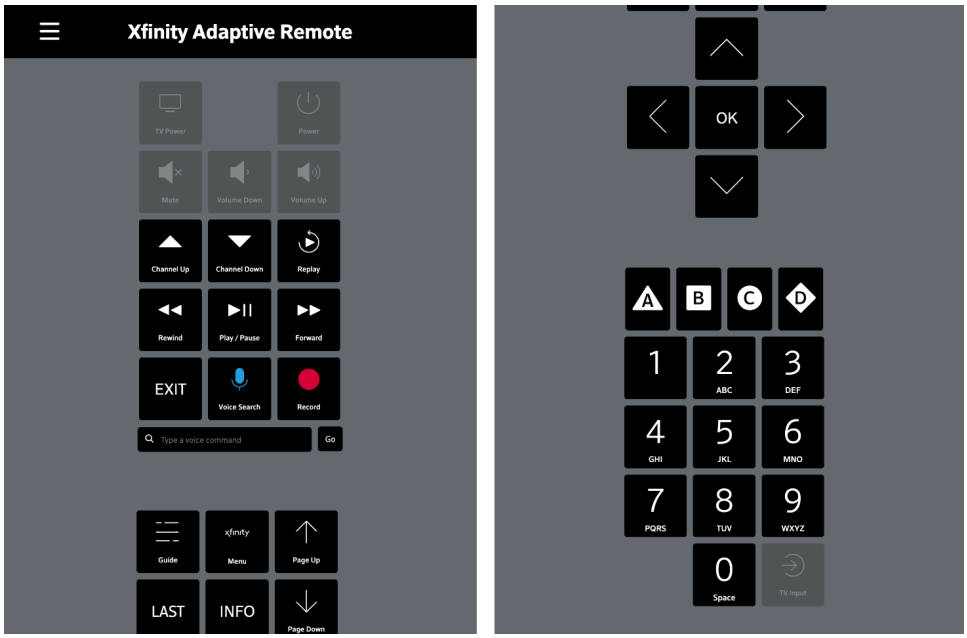

Figure 3 – adaptive remote on iPad (top and bottom scroll)

Tapping buttons on the adaptive remote simulates the press of one of the keys on the physical remote by sending the same vcode to the set-top box that the physical remote would send.

As an alternative to voice control, the adaptive remote has a field to enter a text string that would mimic the desired spoken command. Examples of voice commands include: “NBC”, “Peacock”, “Show me comedy movies”, “Guide”, “Channel up”.

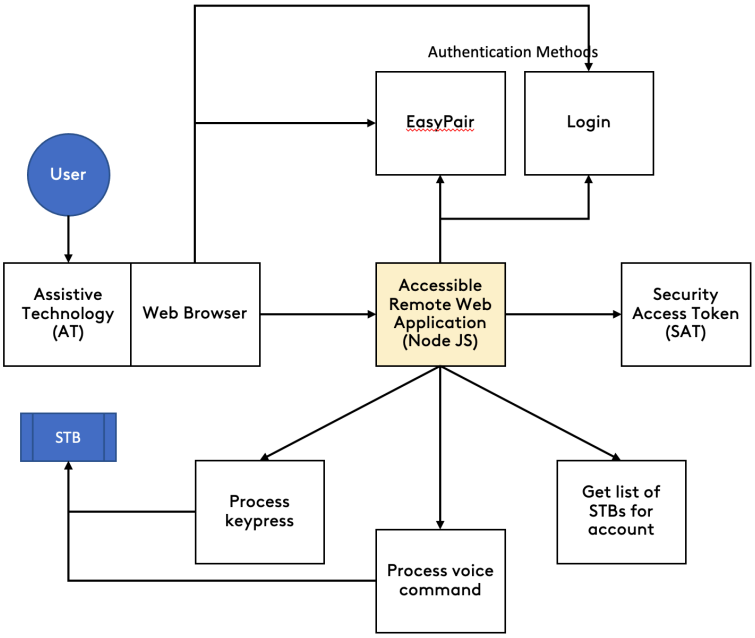

Figure 4 – adaptive remote architecture diagram

5. Companion App

The Adaptive Remote (AccRem) architecture supports “companion applications”. These companion apps use the existing AccRem app for login and TV Box selection and then use an AccRem web service application programming interface (API) developed specifically to support these companion applications. This AccRem architecture enabled the Voiceitt companion application to control the Xfinity set-top box.

Cross-Origin Resource Sharing (CORS)

CORS is an HTTP-header based mechanism that allows a server to indicate any other origins (domain, scheme, or port) than its own from which a browser should permit loading of resources.

Companion IDs

Each run / instance of a companion app must uniquely identify itself by a universally unique identifier (UUID). This is the value that will be implicitly passed via the adaptive remote app when submitting key presses, sending custom text commands, etc.

User Flow

The overall flow for an end user with speech disabilities will typically go something like this:

1. User visits a companion web app, hardware solution, or “fat client” app.
2. The companion app generates a new companion UUID value.
3. The companion app puts up a login button/link with a URL that contains the UUID.

4. Immediately after the user clicks on the button (which opens the AccRem app in another tab or mobile web view), the companion app puts up a "please wait" screen and begins polling using an API pairing endpoint.
5. Once the user logs in and choose her set top box, the AccRem app goes to the companion-success page which shows a message such as "Go back to your Companion App" and the pairing endpoint returns a token value in its response, which causes the companion app to stop polling and move the user to the companion app's "main screen", showing buttons, an input field for entering custom commands to control the set-top box.

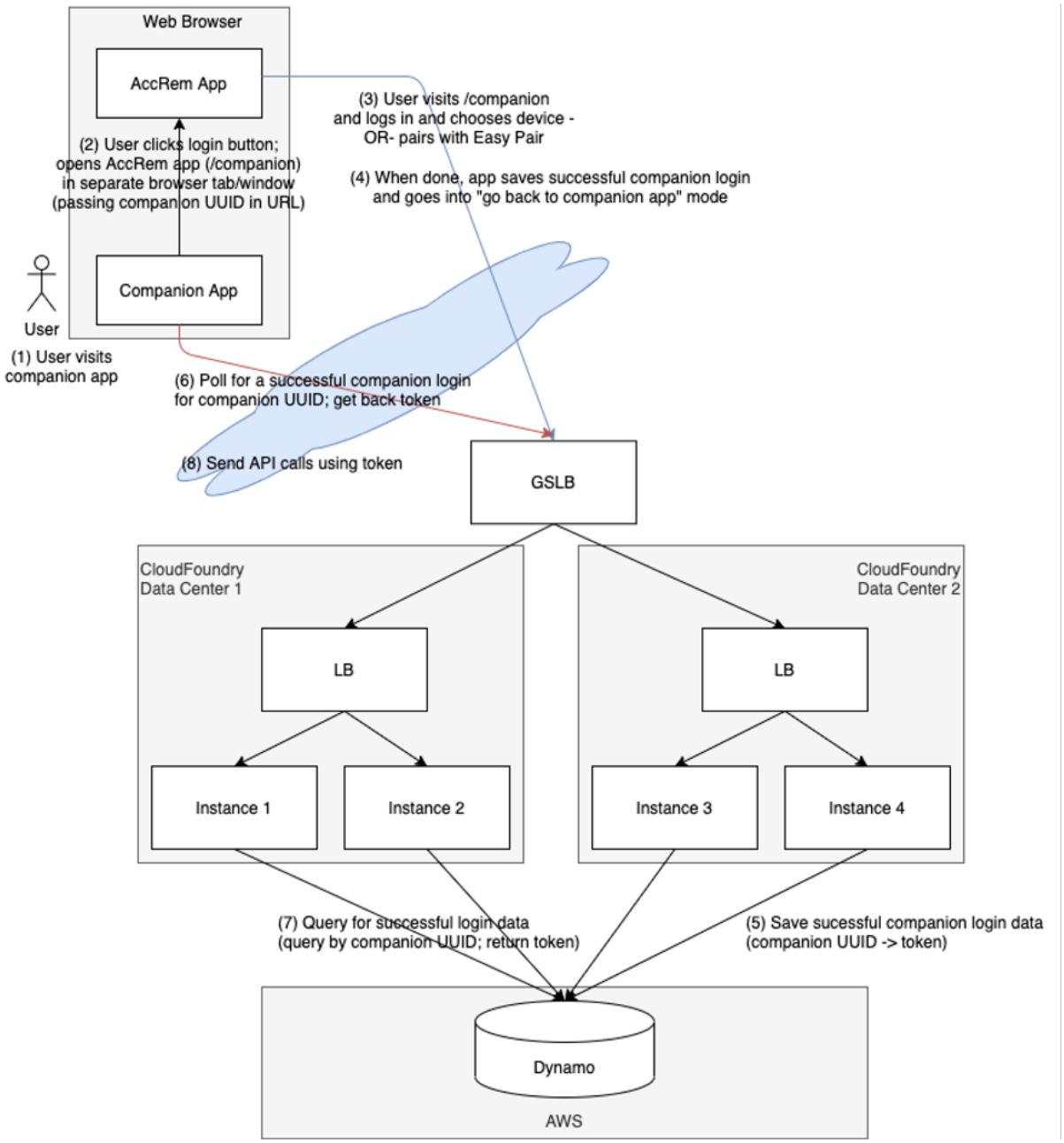

Figure 5 – companion app architecture

6. Integrating the Technologies

Comcast's Companion App was written in NodeJS, therefore Voiceitt translated this to C++ to work with their iOS mobile app. Personal identifiable information (PII) privacy was key in our joint solution.

(We are using our company/product names here to illustrate how we integrated the different technologies while ensuring the privacy of each company's customers. Again, similar solutions can be developed by your organization's product team as well.)

6.1. Customer authentication

To ensure PII privacy for Comcast customers, authentication is done on Xfinity-domain interfaces rendered in the Voiceitt app's web view. No authentication is done through the app itself.

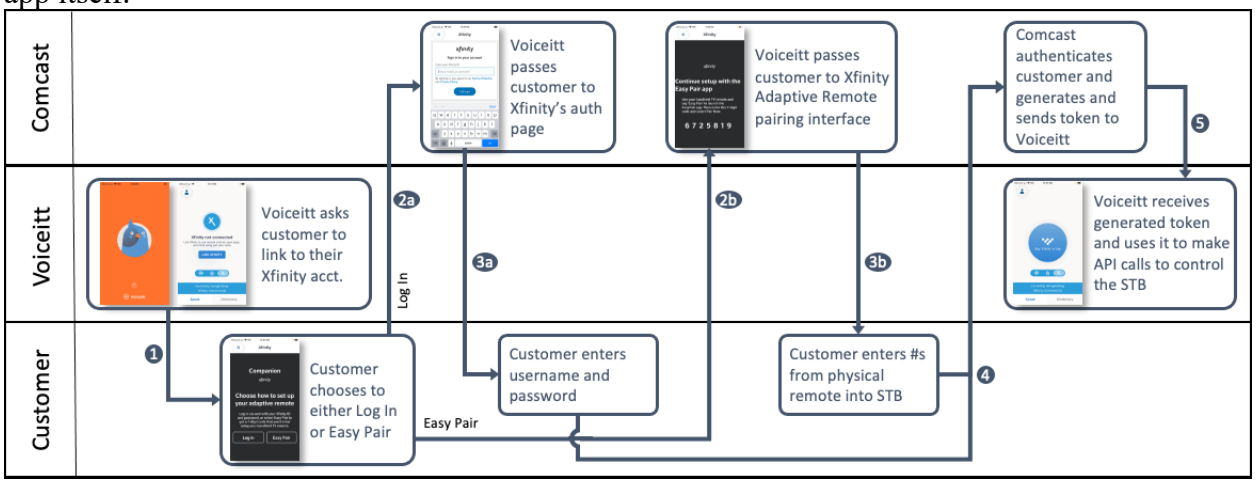

Figure 6 – authentication model

6.2. Voice imprint

To ensure PII privacy for Voiceitt customers, no voice recording is ever shared with Comcast, and moreover is compliant with international data privacy protocols.

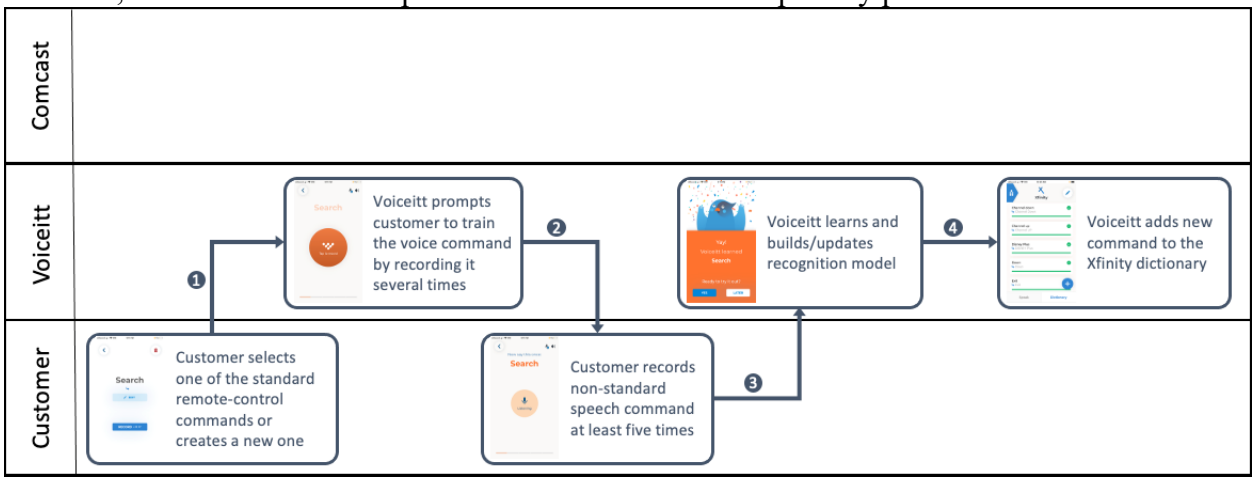

Figure 7 – voice imprint model

6.3. Set-top box control

The cable customer with non-standard speech is now able to control their set-top box with their voice.

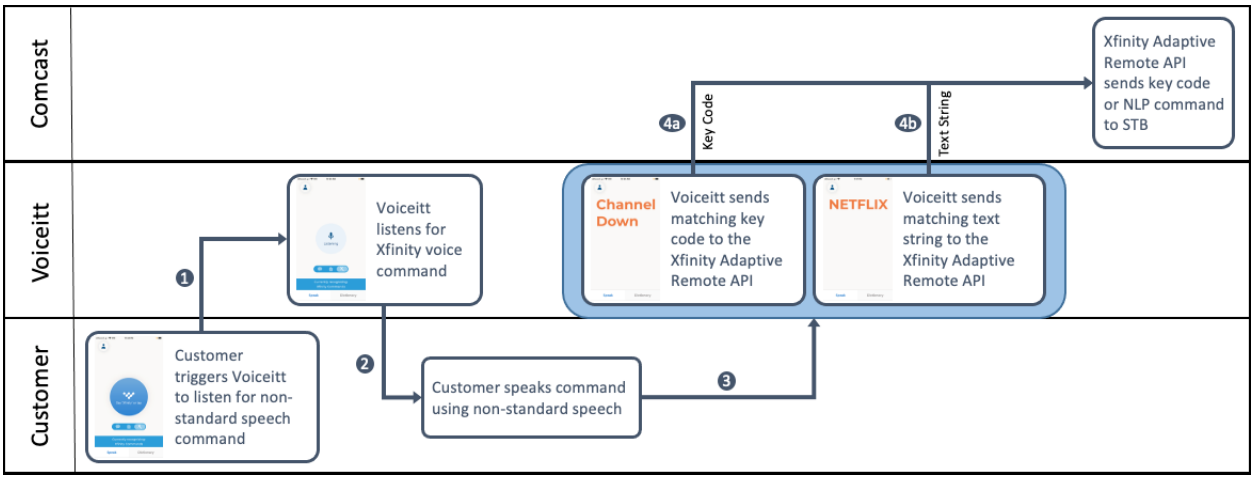

Figure 8 – set-top box control model

6.4. Compatibility with iPhone accessibility options

Any app should be compatible with OS accessibility settings and ATs so that those with disabilities can use it. We ensured all content was readable and in order when using a screen reader. We ensured all actionable items could be reached via finger swipe or AT and were labeled properly. We made sure that all World Wide Web Consortium (W3C) Web Content Accessibility Guidelines (WCAG) were followed at the AA level.

The companion mobile app is also designed and developed with innovative accessible user experience and design so that individuals with motor control impairments, cognitive, and dexterity challenges may access it as independently as possible.

7. Evaluating Impact of Integrated Solution with Customers

Comcast and Voiceitt have collaborated with a specialty nursing care facility in Philadelphia to evaluate the integrated solution described in this paper with end users with dysarthric speech.

The objective of the pilot is to evaluate how the Voiceitt app, which has integrated Xfinity's Adaptive Remote technology, improves independence and quality of life for individuals with dysarthric speech.

In an ongoing pilot, participants with highly atypical speech patterns correlated with cerebral palsy use Voiceitt's customizable speech recognition engine to activate a series of voice commands to their Xfinity X1 set-top box via Voiceitt's consumer application. The participants may not have had prior experience with voice devices or speech recognition; or, they may have previously tried to use these devices but without success. The available voice commands are chosen by each participant, sometimes with the help of a caregiver. The user calibrates the system by recording their voice, following prompts on the screen of their mobile device.

The pilot, now ongoing, will include input from participants, their caregivers and support professionals. Recognition accuracy and daily usage is measured through the companion

application. Impact on customer satisfaction, engagement, and usage, as well as quality of life and independence will be evaluated through a series of interviews and questionnaires with participants, facility administrators, and their daily support professionals.

8. Conclusion

The opportunity to give customers with non-standard speech (especially those whose speech is impaired due to neurological or physical disability) the ability to use their natural voices to control their entertainment system returns to them a sense of independence that offers fewer steps, speedier navigation, less mental and physical effort, and greater overall enjoyment of these offerings.

By integrating two solutions via a companion app bridge, we may serve not only those in the disability community with non-standard speech, but also those with accents, age-related tonal changes, etc. In short, while accessible solutions are necessary for some, they can be helpful for everyone.

We would like to acknowledge Theresa Murzyn and Mike Fine at Comcast. Theresa's user research on how those with ALS and Spinal Injury use media and home entertainment has been invaluable. Mike Fine's assistance in understanding the architecture behind the adaptive remote and companion app APIs has been vital to this technical paper. We would like to give special thanks to our partners at the facilities who provide help in recruiting and supporting individuals with speech disabilities participating in this collaborative pilot.

As our joint pilot progresses, further input from customers with disabilities will inform refinements to the technological approach described here, which will make our solution even more impactful and effective.

Definitions and Abbreviations

ALS	amyotrophic lateral sclerosis, also known as Lou Gehrig's disease, is a progressive neuro-degenerative disease that affects the brain and spinal cord.
Aphasia	The inability to understand what is being said, find the necessary word for something, or formulate sentences due to damage in the brain, often from a stroke or accident.
API	Application Programming Interface
ASR	Automatic Speech Recognition
Dysarthria	A speech disorder caused by either muscle weakness or the inability to control speech muscles due to brain damage.
Content foraging	entertainment system navigation and searching techniques to find content via direct retrieval or orienteering
Easy Pair	A method to connect a remote control to a set-top box by typing the numbers shown on the set-top box interface using the keypad on the remote to be paired.
IoT	Internet of Things
ML	Machine Learning
NLP	Natural language processing
PII	personally identifiable information such as name, address, streaming content history, etc.
RDK	Reference Design Kit (https://rdkcentral.com). RDK is a fully modular, portable, and customizable open-source software solution that standardizes core functions used in video, broadband, and IoT devices.
RDK-V	Reference Design Kit for Video.
STB	Set-top box
Speech synthesis	Artificial production of human speech by computer or speech synthesizer.
UUID	universally unique identifier. A 128-bit alpha-numeric to identify a person, peripheral, etc. without PII (personally identifiable information).
Vcode	Video code. The code sent to the set-top box when a remote-control button is pressed.
Voiceitt	Voice technology startup that has developed automatic speech recognition for non-standard speech.
W3C	World Wide Web Consortium
WCAG	Web Content Accessibility Guidelines. Guidelines written by the W3C's Web Accessibility Initiative directing designers and developers of web applications on accessibility requirements and standards. There are three levels of compliance: A, AA, AAA.
Wernicke's Aphasia	Seemingly fluent speech which is made up of unrelated words (schizophasia, often termed "word salad") or even non-words.

Bibliography & References

Accessible Remote Technical Information, 2021. Mike Fine, Principal Software Architect, Entertainment Experiences, Comcast

Companion App Developer's Guide, 2021. Mike Fine, Principal Software Architect, Entertainment Experiences, Comcast; Adina Halter, Sr. Principal Software Architect, Accessibility Innovations, Comcast

Introducing the New X1 Voice Remote, Dec 11, 2017. Jonathan Palmatier, Comcast

Comcast Wins Emmy Award for X1 Voice Remote Technology, Aug 29, 2017. Comcast

Voiceitt Makes Alexa Accessible for People with Disabilities, PR Newswire

Xfinity Adaptive Remote for Accessibility Audiences, May 19, 2022. Theresa Murzyn, Ph.D., Lead UX Researcher, User Research, Comcast

