Simulating Humans at Scale to Evaluate Voice Interfaces for TVs: the Round-Trip System at Comcast

Breck Baldwin
Lauren Reese
Liming Zhang
Taylor Cassidy
Michael Pereira
Craig Murray
Kishorekumar SundaraRajan
Yidnekachew Endale
Pramod Kadagattor
Paul Wolfe
Tony Braskich
Alice Somers
Donte Jiggetts
Adam Sloan
Esther Vaturi
Crystal Pender
Ferhan Ture
frederick_baldwin@comcast.com
ferhan_ture@comcast.com
Comcast Applied AI
Washington, DC, USA

ABSTRACT

Evaluating large-scale customer-facing voice interfaces involves a variety of challenges, such as data privacy, fairness or unintended bias, and the cost of human labor. Comcast's Xfinity Voice Remote is one such voice interface aimed at users looking to discover content on their TVs. The artificial intelligence (AI) behind the voice remote currently powers multiple voice interfaces, serving tens of millions of requests every day, from users across the globe.

In this talk, we introduce a novel *Round-Trip* system we have built to evaluate the AI serving these voice interfaces in a semi-automated manner, providing a robust and cheap alternative to traditional quality assurance methods. We discuss five specific challenges we have encountered in *Round-Trip* and describe our solutions in detail.

KEYWORDS

voice interfaces, speech, evaluation, nlp

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02.
https://doi.org/10.1145/3539597.3575787

ACM Reference Format:
Breck Baldwin, Lauren Reese, Liming Zhang, Taylor Cassidy, Michael Pereira, Craig Murray, Kishorekumar SundaraRajan, Yidnekachew Endale, Pramod Kadagattor, Paul Wolfe, Tony Braskich, Alice Somers, Donte Jiggetts, Adam Sloan, Esther Vaturi, Crystal Pender, and Ferhan Ture. 2023. Simulating Humans at Scale to Evaluate Voice Interfaces for TVs: the Round-Trip System at Comcast. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3539597.3575787

1 INTRODUCTION

In the last decade, voice-enabled TV systems (e.g., Xfinity Voice Remote, Android TV, Roku TV, etc.) have offered customers the option to freely speak into a remote or hands-free device to find their favorite movies or series, change settings, or tune to a channel. More and more, voice is becoming the default interface for interacting with a TV and the expectation is that it works for everyone all the time, not just as a novelty. For example, Xfinity X1 promotes its voice remote with the motto "Say it, see it".

Evaluating customer-facing voice interfaces benefits significantly from simulated data in that it side-steps data privacy concerns, addresses bias via uniform evaluation of under-represented populations and does all this at considerably lower cost than human generated data or human annotated data. However, the generated data should be correct (e.g. generate very few incorrect inputs) and

the collected data should be greater than or equal to the diversity (entropy) of user inputs seen in production. Finally, it should scale efficiently in both execution and human attention.

Comcast's Round-Trip system presents robust and novel solutions to these issues in an easy-to-use interface shareable from the CTO's office to engineers fixing bugs with many classes of users in between.

2 ROUND-TRIP SYSTEM AT A GLANCE

Round-Trip is a semi-automated evaluation system designed to test a voice interface broadly, by simulating all possible requests (e.g., spoken commands into the voice remote) and validating the correctness or relevance of the action (e.g., display a specific content page on TV). For each program in the catalog, it simulates all possible requests for that program, then automatically classifies the system action into one of few easy-to-interpret pre-defined categories. Based on the classification over an entire catalog, it computes and aggregates metrics that can be tracked over time. In all of this, it provides a user interface that can serve both high-level senior management and hands-on editorial team members, software developers, scientists, as well as product owners.

Round-Trip contrasts with what can be called a whack-a-mole approach that is bound to be focused on high-frequency cases. It serves an entire team responsible for building the AI for voice interfaces and allows them to understand where it is succeeding and where it is failing, in a robust and reproducible manner.

3 CHALLENGES IN EVALUATING A VOICE INTERFACE FOR THE TV

We present the five biggest challenges in evaluating voice interfaces for the TV, using *Round-Trip*:

(1) **Catalog size and complexity:** Typical on-demand program catalogs contain hundreds of thousands of entries, the contents of which shifting dynamically due to programs going in and out of availability. On top of that, there are dozens of streaming apps that offer content of their own, some overlapping and some exclusive. We discuss this complex landscape and the non-trivial engineering and infrastructure scalability problems it causes in *Round-Trip*.
(2) **Query generation:** *Round-Trip* aims to generate realistic text queries with as much diversity as seen in the actual voice requests by customers. Curated templates or heuristics are useful but require manual labor, defeating the purpose of an automated testing system. Language models can be trained based on historical queries but this might perpetuate any harmful bias in the data. We discuss a variety solutions, including hybrid methods that bring the best of both worlds.
(3) **Speech synthesis:** Large-scale voice interfaces are used by people with different speech patterns. Effectively simulating a voice request in *Round-Trip* requires synthesizing speech samples that can ideally represent the diversity of the entire user base. We discuss how text-to-speech models can be trained based on open-sourced data (e.g., LibriTTS), sampled audios from users, or audios donated internally.
(4) **Classification of errors:** *Round-Trip* depends on automatically classifying when the AI fails and when it succeeds, based on the user experience it creates. To ensure a level of reliability, this needs to be done with minimal false positives. System logs are useful in determining the action taken by the AI, but add-on app logs (e.g., Netflix app logs on X1) are not available to the developers of the interface, making this an incomplete solution. Ideally, the classification method should "see" what the user sees and decide accordingly. We discuss deterministic methods based on logs, when available, as well as more complex computer vision methods (e.g., optical character recognition on screenshot image).
(5) **Identifying the root cause:** Ultimately, the intention behind evaluating a voice interface is to fix the cases that are failing, as much as possible. While *Round-Trip* pro-actively identifies these cases, identifying the root cause needs human input. We discuss a process in which editorial staff, product leads, developers, and scientists all work together to review cases flagged by *Round-Trip*, document the root causes and create tasks to fix them, considering speed as well as coverage.

4 CONCLUSIONS

Voice interfaces are becoming ubiquitous as the primary way users interact with their TVs, whether they are watching news on a traditional channel, bingeing the latest hit show on their favorite streaming app, or looking for the picture settings. *Round-Trip* is a novel evaluation system developed at Comcast that is used to test one of the most-used voice interfaces in the world, the Xfinity X1 voice remote, while respecting our customers' privacy and staying conscious about machine bias. As we continue to invest into this new form of quality assurance, we hope to share our learnings and initiate an open conversation that can benefit the broader community.

5 ABOUT US

Company: In 2012, Comcast introduced the world's first broad coverage voice remote control for approximately 20 million cable TV subscribers in the US. To recognize this, Comcast was awarded an Emmy Award for Technology and Engineering in 2018. Since then, Comcast has continued to be a pioneer in voice-enabled content discovery experiences and has expanded the use of its in-house technology globally.

Speaker: Ferhan Ture is a Technical Fellow focused on Comcast's natural language processing strategy and innovation. He is a member of the Comcast Applied AI team, which invents the technological foundations for the Xfinity experiences of the future. Before joining Comcast in 2015, Ferhan graduated from the PhD program of the Department of Computer Science at University of Maryland in 2013, where he defended his thesis on machine translation and cross-language information retrieval.