Bespoken

opearlo

# CASE STUDY

## Opearlo

Bespoken partnered with Opearlo to work on their popular quiz skill Guess My Name. Opearlo, a Y-Combinator-backed company, makes some of the most popular gaming and productivity skills for Alexa. Opearlo and Bespoken have collaborated for some time and saw this as a great opportunity to showcase the benefits of Usability Performance Testing to help make a great skill even better.

# THE CHALLENGE

Bespoken

Guess My Name is an entertaining quiz skill in which the user is given clues as to the identity of a person or thing, and their job is to guess who it is. It's fun and engaging gameplay has made it extremely popular with users. But with hundreds of possible answers to guess, the speech recognition (ASR) and natural language understanding (NLU) for the skill can be challenging – for Alexa, handling complex, custom slot values is a struggle.

"Guess my name is a simple skill with complex Speech Recognition Issues"

These challenges are reflected in the reviews of the Guess My Name skill – here are some examples of recent reviews:

MaryM

★★★★☆ **Needs some fine-tuning, but fun.**

January 27, 2019

Fun game, but it would often misunderstand my response which was correct, but I would be told it was wrong.

Grayson59

★★★☆☆ **Just makes me mad**

January 5, 2019

First of all the games not bad, what makes me mad is after I answer a question(correctly),
It tells me my answer is wrong
then announces the "correct "
answer, which is the answer I
gave and was told it was wrong..
go figure...

7 people found this helpful

Oscar Merry, co-founder, and CTO of Opearlo, and Alexa Champion said this on the challenge of tuning an interaction model for Alexa:

Bespoken to the Rescue!

"There are very few tools available for handling this. What we are typically left with when we make adjustments to our interaction model is submitting a new version of the skill, and then closely monitoring the reviews for issues reported by users. It's a user-unfriendly and unreliable way to diagnose speech recognition problems."

# THE SOLUTION

Bespoken

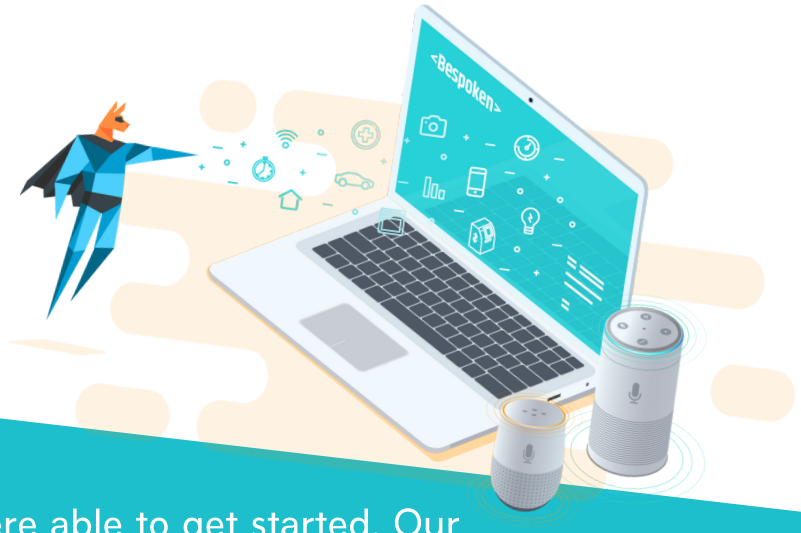
## A better way to diagnose and fix issues

Leveraging our Usability Performance Testing, we ran an initial set of tests against the skill. The tests are easy to set up – we worked with Oscar to identify:

1 **What intents and slots do you want to test?**
In this case, the key intent was the one where the user guesses the answer.

2 **What phrases do you want to test?**
Coming up with phrase variations is important for testing – for example, when guessing a name, users might say: “You are a llama” or “the answer is llama” or even just “llama”. The phrasing affects how well it is recognized by Alexa.

3 **What type of speakers do you want to test with?**
Just as important as phrasing is who the user is – we use text-to-speech from Amazon Polly and Google TTS to create different voices – this allows us to emulate different genders, ages as well as even accents for our speakers.




With this basic information, we were able to get started. Our process for improving the skill was straightforward:

- Run initial tests to get a baseline on performance
- Iterate on the interaction model and the code based on the results
- Run additional tests to see the impact on performance

Using this approach, across a few iterations and a couple of weeks, we were able to greatly improve the skill’s performance.

# THE RESULTS

Our initial tests involved checking thousands of utterances and slot values tested, with a variety of variations across speakers and phrasings. Here is an example of an utterance that was misunderstood:

**Test Details**

| | | | | | |
|---|---|---|---|---|---|
| Name | GuessMyName3 | Intent | CaptureQuestionAnswer | Slot | Answer |
| User | jpk@bespoken.io | Devices | 2 | | |

**Job Details**

| | | | | | |
|---|---|---|---|---|---|
| Start Time | N/A | End Time | N/A | | |
| Success | 6427 | Count | 6841 | Percentage | 93.95% |

**Search** scarf

Before Testing 4/7

| | | | | | | |
|---|---|---|---|---|---|---|
| scarf | FAILURE | a scarf | default | scar | scar | Matched wrong slot: scar Raw value: scar |
| scarf | SUCCESS | the scarf | default | scarf | scarf | |
| scarf | SUCCESS | answer is scarf | default | scarf | scarf | |
| scarf | SUCCESS | is it scarf | default | scarf | scarf | |
| scarf | FAILURE | you are scarf | default | scar | scar | Matched wrong slot: scar Raw value: scar |
| scarf | IGNORE | i am scarf | default | undefined | undefined | Unknown - check the utterance history |
| scarf | SUCCESS | you are scarf | en-US-Wavenet-F | scarf | scarf | |

Results after updating the interaction model to Bespoken's

## USABILITY PERFORMANCE TESTING

After Testing 7/7

| Phrase | Result | Utterance | Voice | Transcript | |
|---|---|---|---|---|---|
| scarf | 7/7 | | | | |
| scarf | SUCCESS | a scarf | default | scarf | scarf |
| scarf | SUCCESS | the scarf | default | scarf | scarf |
| scarf | SUCCESS | answer is scarf | default | scarf | scarf |
| scarf | SUCCESS | is it scarf | default | scarf | scarf |
| scarf | SUCCESS | you are scarf | default | scarf | scarf |
| scarf | SUCCESS | i am scarf | default | scarf | scarf |
| scarf | SUCCESS | you are scarf | en-US-Wavenet-F | scarf | scarf |

Perfect! Overall, our initial results were okay – about 94% of the time, the speech was correctly understood. This might even sound good. But if you step back and think about it, it means more than 5% of the time, the user is NOT being understood correctly. That is a very frustrating experience and leads to the bad reviews. Put it this way – if more than 5% of the time a user clicked a button in your app, it did precisely the wrong thing – how would users feel about it? Would they continue to use the app? And what would they think about the quality of your software? It's a major usability issue, one that voice developers need to proactively address.

Prior to the changes, we saw more than 26% of the reviews referenced understanding and comprehension issues and the average rating was 4.4 stars. With the new version launched, no single user had any complain about understanding and the average review has been 4.7 stars, that's a significant benefit for the users of Guess My Name, as well as for Opearlo. Great reviews mean more users, happier users, and more prominent positioning in the skill store.

# GET STARTED NOW!

Bespoken

> *We worked closely with Opearlo to analyze these results, and Oscar and his team made improvements to the interaction model and the code. After a couple more iterations, we were able to get the acceptance rate to >99%. This means less than a 1% error rate – a more than 80% decrease in errors. This had a huge impact on users, which translated to better reviews.*

★★★★★

## Launch your voice app with confidence!

Is recognizing and understanding your users an issue for you and your team? Now there's a software solution that can help.  Contact us and we'll be happy to get you set up in no time!

To get started just follow these 3 simple steps:

1. Tell us what you want to test:
   Sequence (one-shot, in-session, etc.)
   Intent and/or slot value to testScenarios – types of speakers, phrases, etc.
2. Send us the interaction model
3. Get the results

Email us at: sales@bespoken.io and get a FREE DEMO!