What You Should Know About Choosing Automated Speech Recognition for Your Contact Center

"I'm sorry, I didn't understand that." We've all experienced the frustration of getting that response when interacting with an IVR that uses automated speech recognition (ASR) (also known as speech-to-text) to interact with customers. Yet, some companies are starting to tout amazingly high accuracy levels for their ASR solutions.

Does this mean the problem of machines understanding human conversation is solved? Can contact centers now rely on ASR to transcribe and harvest the vast amounts of data from customer conversations with agents?

The answer depends on the ASR. Human-to-human conversation remains one of the most difficult challenges within the field of artificial intelligence (AI). General-purpose ASR systems haven't suddenly resolved all the issues with speech recognition.

Those reports of high levels of accuracy are often based on data sets that are much easier for machines to understand than typical contact center audio. They aren't being tested against recordings that have poor audio quality, background noise, technical jargon, multiple accents, and other problematic issues common in the contact center.

That's why you need to look beyond generalized benchmarks when you're evaluating ASR capabilities for your contact center.

Explaining the high accuracy claims

The most frequently cited ASR benchmark is based on a public dataset known as the Switchboard Corpus, a corpus of telephone calls collected by Texas Instruments from 1990 to 1991. It's important to understand why accuracy using this data does not translate into high accuracy on audio from modern contact centers:

LANDLINE RECORDINGS

The Switchboard Corpus audio data was collected in the early 1990s, recorded using landline handsets that offered low-noise audio delivered from a stationary location within a generally quiet room. In modern contact centers, customer audio frequently originates from cell phones and wireless headsets and is delivered from a wide variety of noisy mobile circumstances. Audio in the Switchboard Corpus is also normalized for amplitude between the two speakers; but in contact centers there is often a big difference between the quality and volume of the audio for the customer compared to the agent.

NATIVE SPEAKERS

Participants in the Switchboard Corpus data collection were overwhelmingly native U.S. English speakers, which does not reflect the distribution of accents encountered in today's international contact centers.

GENERAL CONVERSATION

The speakers in the Switchboard study were asked to have a conversation with a random stranger about a topic from a predefined list of general-knowledge topics, such as gardening. The result is that the language used throughout the corpus is very general and does not cover any of the obscure jargon specific to a company or industry that frequently comes up in real-life contact center scenarios.

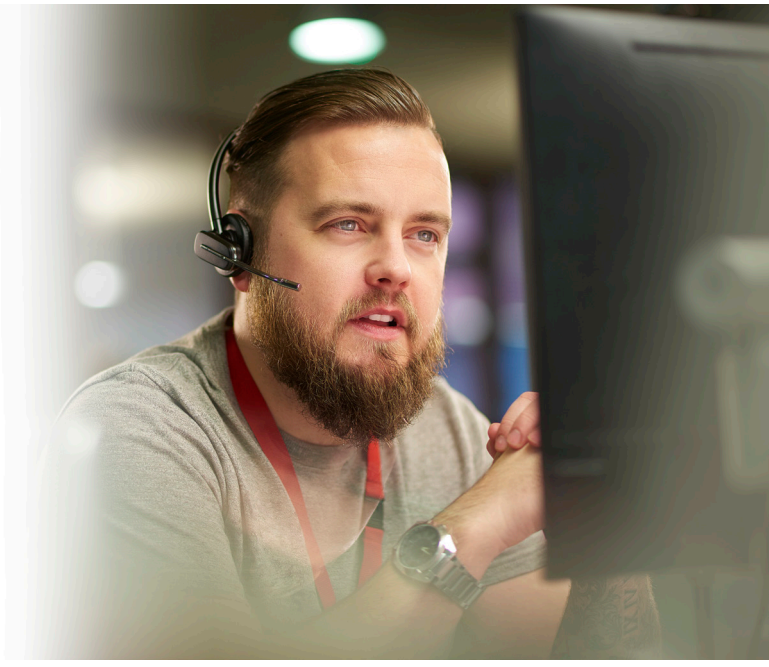
Because they are trained on a dataset with limited complexities, general-purpose ASRs that perform well on the Switchboard Corpus don't perform nearly as well in real-life contact center scenarios.

Benchmarking with real-life contact center data

At Uniphore, it has never been our goal to develop a general-purpose ASR that performs well on the Switchboard Corpus dataset. Instead, we focus on delivering an ASR with exceptional performance in real-life contact center scenarios.

We recently tested our ASR using real-life contact center data. We compared our ASR's accuracy with a popular ASR that reports high accuracy as benchmarked using the Switchboard Corpus. Each ASR was used to evaluate 265 minutes of live conversational data from U.S. contact centers, with accuracy reported as the inverse of word error rate (WER).

Uniphore achieved 80% accuracy compared to the competitor's 63%. This was even after we invested significant work in improving the competitor's results through efforts such as ensuring that the normalizations made by the ASR matched those of our test set references.



Improving ASR performance

The reason Uniphore beats the competition is that our ASR and the other capabilities of our conversational AI platform were developed specifically for contact centers.

For example, one way we improve performance is by treating the customer and agent side of calls differently when we train our models, optimizing for the variations that are common to each side of the call. We also customize the ASR models we develop for each of our clients, using client-specific data to ensure that the range of acoustic scenarios and accents, as well as company-specific vocabulary, are thoroughly represented in the model's training.

Optimizing beyond word-for-word accuracy

Uniphore's hybrid neural ASR engine is trained in a unique way that maximizes performance "downstream" in the Uniphore platform. One problem with typical industry ASR benchmarks is that accuracy is usually given as WER, which treats all words in a conversation as equally important – even filler words such as "the" or "uh" or something similarly insignificant to the goal of the system using the ASR transcript.

When Uniphore studied the correlation between ASR WER and the metric that really matters to contact centers — concept accuracy in the output of the conversational AI system — we found that an ASR optimization objective that treats every word equally does not maximize the accuracy of the final output.

Because the entire stack of our conversational AI platform (from ASR to final output) is developed and owned by Uniphore, we can train our technology using an objective that optimizes for the best possible output in the final system, rather than just for the word-by-word accuracy of the ASR transcript. Most other ASR developers can't optimize for downstream performance.

Improving outcomes with state-of-the-art machine learning techniques

Another reason Uniphore is an industry-leading solution for contact centers is the way we build our models and train our ASR for real-world situations:

Multilingual acoustic feature representations. This approach allows us to tune the acoustic model with small amounts of training data, giving us the ability to ramp up model development for a new language or channel condition quickly while providing a robust acoustic model.

Domain-specific language models. We use a fusion of unsupervised and semi-supervised modeling approaches to tune the language model for a specific client using a minimal amount of human-transcribed, domain-specific data.

Semi-supervised learning. A unique challenge in contact center deployments is the limited availability of transcribed data to train models using traditional supervised methods. However, a vast amount of raw data passes through the speech engine in any deployment. We apply semi-supervised learning methods to tune models on these large sets of data, improving model accuracies consistently without laborious and time-consuming human transcription. This approach is particularly successful for low-resource languages like Vietnamese.

Automatic data augmentation for noise robustness. A contact center ASR needs to handle a wide variety of background and channel noise conditions. We infuse a variety of noise sources into our model training pipeline,

thereby developing noise robustness as part of model building. Unlike generic speech engines, we infuse noise and channel conditions specific to contact center audio channels.

Conclusion

While some generalized ASR systems can achieve high accuracy on certain data sets, only an ASR that's part of a conversational AI platform designed specifically for contact centers can deliver maximum business value within the shortest amount of time. Uniphore is leading the industry in harnessing the power of the human voice to transform the contact center and the customer experience.

uniphore

www.uniphore.com