Speech-to-Text Accuracy in Contact Center Technology

A Guide to Assessing and Improving ASR Accuracy

Speech technologies have been marketed to contact centers for nearly two decades. Early speech-to-text (STT) products were slow, inaccurate, and expensive. Recent breakthroughs in machine learning and artificial intelligence (AI) have vastly improved STT accuracy and speed while other advancements have decreased implementation effort and cost.

Today, contact centers are using highly accurate automated speech recognition (ASR) technologies to gain an advantage over competitors, with AI-driven speech analytics that help organizations better understand caller needs, provide superior service faster than ever, and boost overall operational efficiency.

Why STT Accuracy Matters in Contact Centers

As modern technology and data analysis have made personalized B2C interactions the norm, customers have come to expect more from product and service organizations. Customer experience now serves as a key brand differentiator, with Gartner reporting that 89% of companies compete primarily based on customer experience. Positive interactions drive loyalty and advocacy, while negative interactions drive customers out the door. In contact centers, high accuracy, real-time ASR technologies help brands deliver on the promise of stellar customer experiences through:

- Live agent assist and next best action guidance
- Instant performance feedback, call summaries and evaluations to help agents improve
- Tools to help supervisors give targeted real-time coaching and support
- Process automation, including routine calculations and data entry
- Identification of at-risk customers and product offer or upsell opportunities
- Real-time business and customer analytics

These offerings reduce training time, helping new agents learn fast, and ensure every agent is continuously improving. Agents also require less manual supervision – something especially valuable today as remote and dispersed workforces are on the rise. These time savings help supervisors and managers stay focused on the most pressing and strategic issues. Additionally, managers receive valuable data to support process improvement and team efficiency. Together, these added benefits to the contact center give customers the seamless experiences they expect, where problems and questions are resolved quickly and accurately.

Key Aspects of STT Accuracy

To realize such benefits, contact centers need ASR engines that produce highly accurate text. There are multiple aspects of STT accuracy to consider.

Overall accuracy:

Overall accuracy is based on the number of correct words within a text. For example, a 100-word transcript with 95 correctly transcribed words has 95% overall accuracy. This measure is most important for people reading text. For call analytics, overall accuracy is notable but not the most critical metric. With the volume and repetition of words spoken on most calls, analytics programs can determine sentiment and meaning even with occasional errors. Calls also include many "filler" words such as a, the, of and for where errors have little to no impact on AI algorithms.

Specific word accuracy:

More relevant for analytics purposes is specific word accuracy, or the accuracy of high-value words like company, product and brand names as well as industry-specific terms. It is imperative that an ASR engine recognize the words that matter most to the business and that are used to trigger agent responses, actions, or processes.

Diarization accuracy:

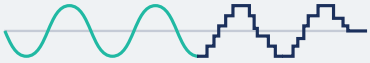
Another factor is diarization accuracy. Diarization is the algorithm used to separate distinct speakers in calls recorded onto a single audio channel. For analytics systems to function effectively and deliver accurate insights, an ASR engine must know who is speaking when. Diarization is not required when each speaker is recorded on their own audio channel.

ASR engine tuning:

ASR engines only recognize words in their standard dictionary, which is unlikely to include product terms, brand names or industry jargon. To improve overall and specific word accuracy, custom tuning can be done to increase recognition for words and phrases commonly used by clients and agents. To reduce time spent on tuning, choose an ASR engine that offers custom languages and pre-defined industry or topic lexicons.

The Importance of Audio Quality

Audio quality plays a big part in every component of STT accuracy. The lower the quality of your audio, the lower your accuracy and the fewer insights you'll acquire. Contact center call audio is a valuable asset that should be treated with care. Recognize that if audio is compressed in size or converted from channel-separated stereo to single-channel mono, the quality is downgraded. In fact, most audio format conversions are lossy, with valuable information about customers and agents permanently lost during each conversion.



To protect your audio quality, follow these best practices:

— Minimize conversions across the audio lifecycle

— Record in stereo to keep the agent and client recordings separate

— Avoid conversions that eliminate channels

— Avoid converting to lower audio bitrate

There are also numerous free tools that can help with audio work and analysis, including:

— **Mediainfo:** For viewing audio container and encoding details

— **Audacity:** For audio playback, frequency analysis and editing

— **Spek:** For spectral analysis (Time-Frequency-Amplitude plots)

— **FFMPEG:** The Swiss-Army knife of audio conversion

How to Test and Measure STT Accuracy

Once you confirm the quality of your audio files, next up is measuring STT accuracy. The 3-step process summarized below uses SCLite, a free tool created by the National Institute of Standards and Technology (NIST), and is a great place to start. The SCLite documentation can provide more details on tool usage and file requirements.

1. Prepare a reference file.

A reference file is a text transcript created manually with high accuracy (>98%) from your sample audio. The audio must be transcribed exactly as spoken. This includes spelling out dates, numbers and URLs such as W W W dot google dot com. The text must not include punctuation, speech fillers (um or ah) or line breaks except between calls. Additionally, each channel needs its own transcript. A reference file can be reused as often as needed for testing.

2. Create a hypothesis file.

Use the ASR engine being tested to generate a transcript for the same sample audio. Because this file will be compared to your reference file, the file formatting should be the same.

3. Measure the accuracy of your hypothesis file.

Use the SCLite comparison tool to produce detailed accuracy reports. These will include information on:

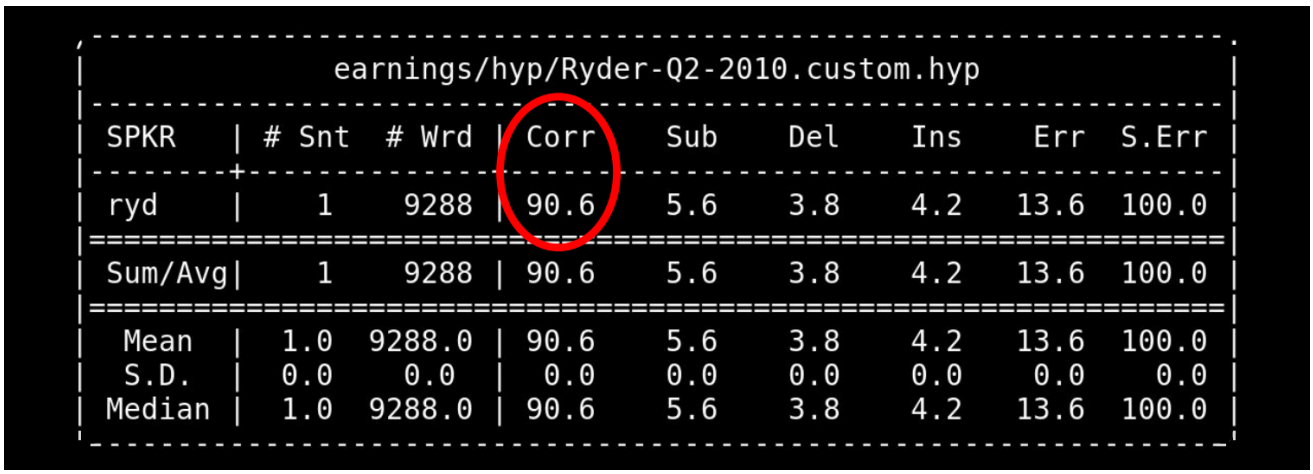
— Total number of errors and their individual locations

— Number of errors by category: substitution, insertion and deletion

— Correctness, measuring the number of correct words in the hypothesis compared to the total number of words in the reference

— Word Error Rate (WER), which is the total number of errors divided by number of words in the reference

To assess specific word accuracy, be sure to review the errors within the text; again, some errors have a bigger business impact than others. If you see false positives due to contractions and spelling variations (like OK and OKAY), you can adjust the reference file to match the variants used in the hypothesis file and rerun the test for more meaningful measurements.



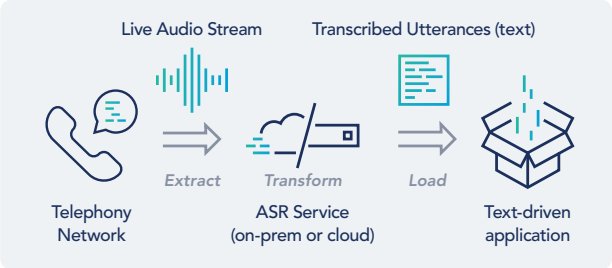
earnings/hyp/Ryder-Q2-2010.custom.hyp								
SPKR	# Snt	# Wrd	Corr	Sub	Del	Ins	Err	S.Err
ryd	1	9288	90.6	5.6	3.8	4.2	13.6	100.0
Sum/Avg	1	9288	90.6	5.6	3.8	4.2	13.6	100.0
Mean	1.0	9288.0	90.6	5.6	3.8	4.2	13.6	100.0
S.D.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median	1.0	9288.0	90.6	5.6	3.8	4.2	13.6	100.0

SCLite Summary Report Example (https://github.com/usnistgov/SCTK)

Additional Aspects Impacting STT Technology Success

While accuracy is critical, it's far from the only thing that influences the success of contact center STT technology solutions. Here are a few additional factors to look for.

Speed: Low STT latency is required for applications like real-time in-call analytics and live agent assist that truly transform contact center operations and caller experiences. For both real-time and post-call processing use-cases, higher throughput (audio hours processed per hour per server) means lower operating costs by getting the same amount of work done with fewer servers.



Completeness: Look for ASR engines with features such as:

— Punctuation, capitalization and number formatting
— Rich metadata to enable advanced analytics (ex: emotion, sentiment and gender)
— Multiple language and dialect support
— Redaction and encryption for security

Simplicity: To get an ASR engine up and running fast, well-documented REST APIs are best. Many settings should be available to enable optimization, but they should be optional.

Flexibility: Look for technology that is scalable and adaptable to changing business demands and tech stacks, ideally offering on-premises, private cloud and pure SaaS offerings as well as supporting batch file and real-time streaming process flows.

Vendor track record: If a vendor has a history of improving accuracy, lowering costs, adding languages and increasing ease of use over time, those trends will likely continue.

An Industry Leader in Accuracy and Speed

The enterprise-grade ASR from Voci Technologies provides best-in-class accuracy, with **a 10% out-of-the-box improvement in STT accuracy** over alternative technologies. Vertical and custom language models boost accuracy for contact centers right away.

Voci's STT technology automatically redacts sensitive information to keep callers safe and secure, while providing true real-time transcription complete with punctuation, speaker separation, gender, emotion and sentiment to help contact centers generate valuable insights.

Voci also delivers ease of integration and the lowest total cost of ownership, with flexible same-day deployments, REST APIs, horizontal scaling and the ability to transcribe 1 audio hour every 10 seconds per server.

Visit vocitec.com or call 412-621-9310 to learn more about how Voci's real-time STT technology can power software solutions that unlock insights from rich voice data and that improve caller experience, agent performance and operational efficiency across contact centers.

Voci Technologies, the leading speech analytics platform provider, enables enterprises to gain actionable insights on their terms from 100% of customer calls. For information, visit www.vocitec.com.