Speech Analytics Transcription Accuracy

Understanding Verint’s speech analytics transcription and categorization accuracy
# Table of Contents

Introduction ..................................................................................................................... 1

Realizing the Speech Transcription Potential ................................................................. 1

Defining Accuracy for Speech Transcription ................................................................. 1

- Word Error Rate (WER) ............................................................................................... 1
- Precision and Recall .................................................................................................... 2
  - Precision (Correctness) ............................................................................................ 2
  - Recall (Completeness) ............................................................................................. 2
- Balancing Precision and Recall ................................................................................... 3
- Perceived Business Accuracy ..................................................................................... 4
- Interpreting Transcription Confidence Scores .............................................................. 4

Tuning Accuracy for Your Unique Data ........................................................................... 5

Going Beyond Transcription Accuracy with Conversational Analytics ....................... 6

- Speaker Labeling ......................................................................................................... 6
- Operational KPIs .......................................................................................................... 6
- Categorization Accuracy .............................................................................................. 6

Detecting Emotion and Sentiment ................................................................................... 6

Powering Speech Analytics with 100 Percent Transcription ......................................... 7
Introduction
Speech analytics transcripts are the bedrock of many predictive and prescriptive efforts undertaken by organizations around the world.

But transcribing a contact center conversation is a deceptively difficult task. Transcription accuracy has reached near-human levels in a simple, single-speaker dictation. However, with the combination of background noise, hold messaging, cross-talk, colloquialisms, poor audio quality, multiple speakers, accents, and more, accurately transcribing a contact center conversation can be a significant challenge.

Despite this challenge, the business-critical uses of speech transcripts make it even more important that the transcripts be highly accurate. How is accuracy actually determined? Even measuring accuracy can be a complex issue, with different measures and balance-points, each of which must be considered in context of the intended use case.

This document provides an overview of the various methods and metrics for measuring transcription accuracy, and how they can be used to assess the overall effectiveness of a speech-to-text transcription strategy. It also presents findings from actual use cases showing how organizations are benefitting from speech transcription.

Realizing the Speech Transcription Potential
Today, most commercially available speech analytics solutions fall into two types: (1) full-text transcription of the spoken word, or (2) phonetics-based output of the sounds heard, rather than the words themselves.

Verint® is a pioneer in speech analytics, commencing research in the late 1990s and launching a commercially available speech analytics solution targeted to the contact center industry in 2003. Building knowledge through scientific rigor, we have found that analyzing unstructured call recordings through full transcription not only yields more accurate results, but also transcripts that are the building blocks for a far more robust understanding of these rich customer conversations. Speech analytics transcripts are now an integral component of our artificial intelligence and automation product offerings, and their accuracy is of paramount importance.

Defining Accuracy for Speech Transcription
Several different metrics are collectively referred to as “transcription accuracy,” with each providing a different dimension of accuracy and impact on transcription results. Let’s take a closer look.

Word Error Rate (WER)
The most commonly-used transcription accuracy metric is Word Error Rate (WER). Word Error Rate is derived from the Levenshtein distance — the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into another, working at the word level.

The example below presents a conversation between an agent and a customer and shows how WER is calculated.

The icon represents the manual transcription, representing the actual words that were spoken on the call. The icon represent the automated transcription, as provided by the transcription engine.
The smaller the WER, the higher the transcription accuracy.

**Precision and Recall**

The WER transcription accuracy metric, although commonly used, is limited in that it does not provide details on the nature of the transcription errors. Metrics such as **precision** (correctness) and **recall** (completeness) fill that gap by implying the impact of the error on the analysis.

**Precision (Correctness)**

Precision indicates how many of the transcribed words are correct. If the transcription engine includes a word that was not spoken in the dialogue, this adversely affects precision. In business applications such as speech transcription, the metric indicates the level of accuracy or reliability of a transcript provided by the solution.

**Recall (Completeness)**

Recall indicates how many of the words spoken appear in the transcript. In the academic world, this is referred to as a measure of **recall** or **detection rate**. From a business value perspective, it is reflective of the degree of completeness of transcribed conversations. If a word is spoken in the dialogue but not included in the transcript, this adversely affects recall rates. For example, a call that is transcribed at 50 percent recall will have only 50 percent of the spoken information contained within it.

The measures of **precision** and **recall** are exemplified below, on the same sample conversation presented above. The correctly transcribed terms are underlined on both the manual transcription and the automated transcription.

Word error rate:

\[
\text{WER} = \frac{N_{\text{substitutions}} + N_{\text{insertions}} + N_{\text{deletions}}}{N_{\text{words}}}
\]

Accuracy equals 1 – WER.

In our example:

\[
\text{WER} = \frac{4 + 2 + 1}{26} = 26.9%
\]

Accuracy = 73.1%
Reviewing this example:

- 26 terms were provided on the manual transcript, with 21 of them correctly transcribed automatically, yielding a recall of 80.8 percent.
- 27 terms were provided by the automated transcription, with 21 of them correct, yielding precision of 77.8 percent.

Balancing Precision and Recall

Now that we understand the available measures of accuracy, we need to determine which of those metrics we should use to represent it. The decision depends on the tradeoff between precision and recall.

Let's examine an example of this tradeoff by looking at a customer retention scenario in which the customer retention team wants to be informed whenever a customer verbalizes an intent to terminate business with the organization.

If the impact of losing each customer is high and the retention attempt cost is low, a solution with high recall would likely be preferred. This is because a solution with high recall would be less apt to omit occurrences of words indicating customer churn.

On the other hand, if the retention team can handle only a limited number of customers per day, a solution tuned to higher precision to identify the customers most at risk or with the highest lifetime value would likely be more effective, even at the cost of lower recall. This is because
a solution with a high degree of precision would be less likely to produce false-positives, indicating that a keyword was said when in fact it was not.

Any solution can be tuned to achieve 100 percent recall at the cost of very low or zero precision, or 100 percent precision but with very low or zero recall. In reality, the right balance between these two measures is dependent on the business use case, but it is important not to go too far to any one extreme, sacrificing precision for recall or vice versa.

Perceived Business Accuracy
Another important consideration when measuring accuracy is whether all words are equally important. This question is significant from a business perspective. Perceived business accuracy suggests that some words and terms associated with key insights are more important than words, such as “have,” “is,” and “the.” Reviewing the examples above, some of the errors may be more important than others. For example, “I” vs. “I’ve” or “question” vs. “questions” may be less important for analysis, while “If” vs. “refunded” may be more significant.

Using data-driven adaptation techniques, Verint can ensure that the transcription engine is focused on the most significant terms that are relevant to businesses. Narrowing the assessment of transcription accuracy on business-relevant terms typically increases accuracy by a few percentage points, but the positive impact to business analytics and insights is typically much more significant.

Interpreting Transcription Confidence Scores
Verint’s transcript includes confidence scores for every transcribed word or term. The confidence score is a measure of reliability of the transcribed word or term over other alternative possibilities. The score values range between 1 and 1000, where a score of 1000 means that the transcription engine is almost certain that the word transcribed is the word spoken, and 1 means the engine has no certainty at all. Verint Speech Analytics™ leverages these confidence scores to exclude noisy elements from analysis.

For example, words with very low confidence score may not appear in the transcript within the call player, even if they may be accurate. These terms still affect various algorithms (including categorization scores) and are included in the transcription export.

Confidence scores for words or terms should be compared only to confidence scores within the same transcription model / language. For example, confidence scores for banking customer service calls handled in English have no correlation to confidence scores for banking customer service calls handled in Portuguese, and certainly no correlation to healthcare billing calls handled in Spanish.

Confidence scores define the probability of an event — in this case, accurate transcription of a word or term. Confidence scores are calculated at the word or term level. On the other hand, accuracy defines the skill of the learning algorithm to predict accurately. It defines the percentage of correct predictions made among all of the predictions made in transcribing a call in its entirety. While there is a strong correlation between confidence scores and WER, they are indeed different.
Tuning Accuracy for Your Unique Data

Verint’s rich linguistic and acoustic models, developed and refined with over a thousand speech analytics deployments, processing over 20 million calls each day in over 60 languages and variants, can offer outstanding, out-of-the-box transcription accuracy.

Still, even within the same industries, individual businesses have unique words, terms, phrases, product names, and legal regulations that are rare in everyday spoken language. To help ensure the success of every customer, Verint has developed methods for quick and impactful adaptation of the base model to each customer’s unique environment. This tuning process can provide even higher transcription accuracy.

The specific language model accuracy levels can be visualized in three key levels:

- **Base Level**: The out-of-the-box language model providing the broad and rich foundation of speech. For Verint, this base layer includes over 60 different language models.
- **Middle Level**: The language and accent customization (LAC) layer that extends the base model by adding terms and phrases that are prominent to the customer’s unique environment, and are learned automatically from a training set provided by the customer. Powered by advanced machine learning capabilities, this layer commonly yields an additional 3 - 10 percent accuracy improvement, depending on how unique the customer environment is compared with the base model.
- **Top Level**: The phonetics boosting layer, which includes fine-tuning of unique business-critical terms and phrases, as defined by the business users. This portion of the model can be dynamically edited and altered by users throughout the customer lifecycle, and may include various elements, such as:
  - Addition of emerging terms, such as new campaigns or product names, new regulations, competitors, or even the names of the company’s senior executives.
  - Correction of incorrectly recognized and/or transcribed terms for continuous accuracy boosting.
  - Balancing of similar-sounding terms, such as “loud” and “allowed.”

This phonetics-boosting layer helps improve the recognition of specific terms and phrases unique to the organization, improving the perceived business accuracy and the relevant categorization accuracy, while the LAC layer improves the overall transcription accuracy in cases where enhanced tuning is needed. As expected, the more feedback and tuning the system receives from its users, the more accurate it becomes.

The actual precision, recall, and WER can vary significantly from environment to environment and from interaction to interaction, and are dependent on many factors as noted above. When applying the appropriate base model, relevant LAC, and phonetic boosting, Verint’s transcription accuracy on contact center audio is frequently unmatched. We continue to invest in innovations and ways to improve transcription accuracy even further.
Going Beyond Transcription Accuracy with Conversational Analytics

While transcription accuracy is obviously extremely important, there are also other factors to consider beyond word accuracy. When analyzing customer conversations, it is important to know who the speaker is, link conversation transcripts to operational KPIs, and accurately categorize the types of conversations taking place. Adding these metrics to the transcription can provide the speech analytics user and data scientist with a significant lift in value and business accuracy.

Speaker Labeling
Speaker labeling provides analysis of the call dynamics between agents and customers, and can offer valuable insight when incorporated into the larger analysis of a key business issue. In some cases, the recording environment is able to provide speaker-related information. However, in recording environments where speaker labeling is not available, Verint’s speech analytics solution is able to provide this information via a patented software-based algorithm. The labeling precision accuracy is typically over 95 percent with over 85 percent recall, providing consistent measures for conversational analysis.

Operational KPIs
Operational use cases of call transcription vary greatly, from cost containment efforts focused on reducing call volume and Average Handle Time (AHT), to customer experience improvement initiatives evaluating the relationship among caller intent, call outcomes, emotion, themes, and relations. Some of the important KPIs contributing to a complete understanding of the root causes of contact center performance can be derived from the transcription call dynamics, such as agent and customer talk time, silence time, and talk-over time. These metrics are an integral part of the Verint speech analytics solution.

Categorization Accuracy
Categorization accuracy is measured by validating the automated call categorization with subjective human annotation. Each call within Verint Speech Analytics is categorized with a relevant score between zero and one, which is reflected in the application with a five-star ranking scale. Leveraging Verint’s unique complete semantic index and semantic intelligence layer, users can apply various capabilities, including guided and predictive search, context visualization, automated themes and relations, and unique syntax operators to help improve categorization precision and recall. For example, adding a rule to look for the word “close” near the word “account” to a customer churn category can significantly improve categorization precision and recall (even if several words between “close” and “account” are mis-transcribed). Leveraging these out-of-the-box functions, many Verint customers have achieved precision levels of above 90 percent in most categories. In a customer test carried out by a Verint Speech Analytics customer, categorization precision was measured at 94.3 percent as compared to human annotators.

Detecting Emotion and Sentiment
As we all know from experience, discussions between contact center agents and customers can sometimes become emotionally charged. Whether an agent understands what a customer is saying—and even feeling—can make the difference between a good customer experience and poor one; between retaining a customer and losing one.

Verint Speech Analytics can help organizations identify when and where emotional events take place and can proactively analyze and report on these conversations.
Sentiment analysis originated with text analytics, where sentiment can be detected through modifiers such as exclamation points, negation expressions, etc. The Verint Speech Analytics platform provides the ability to identify interactions in verbal communication that contain emotional content. Emotion detection includes the analysis of words and phrases used in the conversation. The analysis can leverage call dynamics and speaker separation—for example, focusing on emotional content from the customer where talk-over was present. Interactions that contain such content are grouped into a category that can then be analyzed in conjunction with other categories.

Verint Speech Analytics can identify emotion based on positive or negative attributes associated with different key words and phrases within specific category definitions. It can then detect different types of emotional reactions such as anger, annoyance, excitement, and confusion. Not only can the solution categorize these calls into sentiment type categories, users can also define who the speaker is, and identify the point in the conversation at which these words or phrases are spoken (for instance, calls that ended with negative emotion by the caller).

In over a decade of research and development, we’ve determined that the most precise mechanism to detect emotion is what people say, not how they say it. Pitch, tone, and other acoustic elements can give too many false positives when callers merely raise their voices or speak quickly to be heard, and miss complaints and dissatisfaction expressed while speaking normally.

Powering Speech Analytics with 100 Percent Transcription

The numerous customer conversations captured in your contact center or branch environment represent a wealth of information that can be tapped through accurate call transcription and powerful speech analytics.

Verint’s high degree of transcription and categorization accuracy has enabled leading organizations to build and enhance predictive churn, collections, and customer satisfaction models to provide real-time guidance to contact center agents, automate quality monitoring efforts, detect fraud, better match product/service offerings to customers, design more effective virtual assistants, IVRs and knowledge management systems, and much more. Contact us today to learn how our solutions can benefit your organization and help you simplify, modernize, and automate customer engagement.

Verint. Powering Actionable Intelligence.®

Verint® is a global leader in Actionable Intelligence® solutions with a focus on customer engagement optimization, security intelligence, and fraud, risk and compliance. Today, more than 10,000 organizations in over 180 countries—including over 85 percent of the Fortune 100—count on Verint solutions to make more informed, effective, and timely decisions.