Gridspace

Gridspace Competitive Differentiators

What makes Gridspace Sift Different?

Enterprise speech systems were first commercialized when the best speech technologies in the world could only translate and interpret short, predefined utterances. While these systems could support basic IVR and keyword detection functions, they were too rigid, slow and inaccurate to analyze actual conversations between people -- let alone many thousands of live, simultaneous interactions across multiple sites and domains.

Today, an entirely new approach to speech processing is possible in the enterprise. Thanks to new breakthroughs in artificial intelligence and high-performance computing, it is possible to distill vast amounts of real-time conversational speech audio, not just short utterances, like other enterprise data streams.

Gridspace Sift was built from the ground up for conversational speech. It allows businesses to automatically turn spoken conversations into structured entities, service metrics, classification labels, prediction scores, and similarity rankings -- among other tasks that would otherwise require tremendous human effort. Companies use Gridspace Sift to deeply understand and quickly react to their customers, employees and markets.

This document summarizes many of the unique features and capabilities that distinguish Gridspace Sift from legacy ASR (automatic speech recognition) and speech analytics solutions. Because terminology can differ across companies and industries, a glossary of terms is included in the back. We hope you and your team find this briefing informative and inspiring ahead of your Gridspace Sift evaluation!

Contents

Live Analysis

Every component layer in the Gridspace Sift stack — from telephony to high-level natural language understanding — is designed to operate live. Gridspace Sift provides live transcripts with a fraction of a second of latency. The system also enables live scanner, topics, and grading results as calls unfold. Finally, Gridspace Sift delivers truly live speech analytics, which are computed and updated live with every new conversation, so operation teams can quickly react to new opportunities and emerging issues.

Gridspace Sift directly integrates into a variety of systems over SIP. Moreover, partnerships with a variety of SIP and TDM ecosystem partners make configuring Gridspace Sift for live processing straightforward. While IP-based integrations are common among legacy providers, most other providers batch or entirely avoid sophisticated live analysis. Some legacy systems and cloud solutions even require companies to transfer or manually upload recordings before any analysis can begin.

ASR

Automatic Speech Recognition or speech-to-text systems serve an important function in Gridspace Sift's conversational speech pipeline by delivering clean machine transcripts to natural language understanding components. For decades, computer and algorithmic challenges limited ASR systems to short-form speech transcription tasks. One of the first ASR systems created was limited to recognizing numbers between one and ten. More recent ASR systems, such as those from large cloud platforms, were optimized with VPA (virtual personal assistants) in mind and were never suited for making sense of long utterances.

Gridspace Sift ASR components are different. They were built for human-to-human conversations and, accordingly, trained on in-domain speech audio from human-to-human conversations. Training with in-domain audio and corpi has helped Gridspace achieve best-in-class, in-domain accuracy nearing 97% accuracy on select conversational speech datasets while cloud speech APIs achieve less than 44%. Moreover, Gridspace's training process helps ensure consistent outputs by factoring in how and where people talk. Models are robust to variations in noise, reverberation, microphone, speaker, and dialect.

Not to be confused with ASR, phonetic search vendors approach long-form speech using hybrid techniques that avoid full transcription. Rather than transcribe, these methods perform "phonetic indexing" using simple signal processing algorithms. While these methods require minimal computing resources, they have extremely low accuracy, and do not generate a transcript, preventing downstream machine learning. These systems can only produce static alerts and rudimentary search hits (often with high false positives). In contrast, Gridspace Sift benefits from cutting-edge deep neural network transcription models, which are extremely accurate and robust to noise.

Scanner

Many call analysis tasks require a system to 1) verify if a rule was followed or subject was discussed and 2) extract target information from a conversation. For such tasks, Gridspace Sift provides a powerful and adaptable capability called Scanner. Scanner allows a company to match and extract concepts in conversations without explicitly authoring exhaustive rules and conditions.

Scanner is especially useful for applications related to compliance, agent coaching, CRM data entry, and alerts, in which the target result maybe ambiguous and appear anywhere in a conversation. Scanner can also act as the language for performing semantic searches across historical calls. This enables analysts to query past interactions with the same scripts as live calls.

Here is a quick example: say a company wished to be alerted whenever pricing was discussed, a Scanner query of the form ~'pricing' tells Scanner to look for words or phrases that are semantically related to pricing (the tilde indicates we're looking for approximately 'pricing'). If the user wanted to instead find only upset callers asking about pricing, they could scan for ~'pricing' and '{negative}'. If they wanted to further restrict to callers who then cancelled their account, they could scan for ~'pricing' and '{negative}' then ~'cancel'. This queries would catch a variety of interactions that literally fit the bill!

Scanner also allows analysts to seamlessly blend searches for words or phrasing with generic concepts like emotion, sentiment, or abstractions (like names or dates) as well as some acoustic concepts such as anger or a baby crying, laughter, and music.

Unlike most supervised machine learning systems, Scanner requires no "training data" and simply relies on high level instructions from

users to generate a model on the fly. While some legacy systems can detect single transcribed keywords in transcripts, nobody offers the speech capabilities that Scanner provides. Scanner is the first speech tool that allows fluid natural language descriptions to define a machine learning model. Even in the wider field of text analytics, Scanner is unique to Gridspace.

Call Grading

Call Grading is a class of models that Gridspace generates from labelled calls, designed to look at both the semantic (what was said) and acoustic (how it was said) contents of a call to emulate a human evaluation of the same call. Call Grading models can learn to emulate customer surveys and call center manager QA reports with surprising accuracy.

Call Grading is a highly-specialized blend of machine learning models that has sensitivity to both the content of the discussion as well as speech undertones, such as cadence, tone, and other prosodic features. The Gridspace grading models utilize recent advances in recurrent neural networks, convolutional neural networks, speech features, and our high accuracy ASR (which is one of the inputs to these models).

While other “call data mining” solutions perform rudimentary call classification, these systems are limited by low-quality batch transcripts and hand-designed alerts and keyword detections. Such systems may catch obvious signs a call is going well or poorly (i.e. catching specific, hard-coded good or bad words), but miss high-level patterns, sensitivity to tone and meaning.

Similarity

Call Similarity provides an interface to answer the question, “of all past calls, which were most similar to this conversation?” Similarity, like Call Grading, looks at both the semantic and acoustic content of the call to holistically compare calls across a large set.

One important benefit of Call Similarity is the ability to look up a similar call that has metadata — call center coaching or a summary — and mirror that metadata from the most-similar neighbor. Another use is as a clustering metric that preserves very fine similarities between calls.

While other call and text analytics platforms offer basic clustering (typically using off-the-shelf text clustering methods), these algorithms typically have a weak understanding of fine similarity between calls and hang on a single word or tone of voice.

The Gridspace Call Similarity algorithm is a blend of neural networks and text transformations that were invented and developed at Gridspace, specifically for analyzing large volumes of calls that vary in subtle but meaningful ways. While competing clustering features may provide some insight to the landscape of calls in a contact center, Gridspace Sift’s similarity capability is highly specialized for natural speech and cross-conversation meaning.

Topics

The Gridspace Topics processor can run on every interaction through the Gridspace Sift system. Topics uses a deep neural network to evaluate which words and phrases from a call are uniquely “useful” given the context. Some call analytics tools will present a “word cloud”, which can be a fun way to explore the content of a call. But Gridspace Topics goes beyond frequency and mimic the keywords and phrases a real person would use to describing a conversation.

Another alternative approach to topic generation is classical “Topic Modelling”. This family of techniques can categorize and cluster individual words used in a conversation, but are unable to actually extract important (and often rare and unique) phrases. These models only are capable of giving a rough idea of language used in calls, and the obscure results.

The Gridspace Topics model is designed to form part of a compact, lightweight summary of a conversation. Additionally, Sift computes “Trending Topics”, which can quickly identify emerging topics that are both useful and unique to recent calls.

Search

Indexing conversational speech like one might index text ignores the unique properties of speech and hampers the exploration of large conversational speech datasets.

Most speech search tools approach speech one of two ways: The first is traditional text indexing. If the service generates a full transcript, the text is indexed so that queries can be performed quickly and rigidly. A search returns instantly whether the word “gold” occurred in the transcript.

The second approach is phonetic indexing. Older solutions rely on this technique that cannot fully transcribe and roughly index sound of words. The query “gold” might also return “old”, “fold”, “sold”, “god”, and “told”. This can be useful when false positives are tolerable and resources are constrained.
Pure phonetic indexing is extrem
ely inflexible and inaccurate. It can be useful when false positives are tolerable (a compliance investigation into a small dataset), but is unusable for searching a large dataset of calls with specificity. Rare phrases (needles in a haystack) are intractable with phonetic indexing, but trivial with Gridspace Sift.

Traditional search is only as accurate as the transcribing engine used. Legacy ASR systems in other products, often have error rates 3-5x worse than Gridspace. This translates to higher false positives and higher false negatives. Additionally, the query “gold watch” is unable to find similar phrases like “silver rolex”. Gridspace Sift incorporates Scanner into its search engine, which allows for semantic fluidity that doesn’t exist in other solutions.

Recording

Recording is a simple operation with Gridspace Sift and fully controllable via dashboard and API. All audio is by default stored and when possible, as speaker-separated audio streams. Additionally, audio is stored in high-fidelity format so playback is clear and so new models can be run on previously-recorded audio. All audio recordings are stored on AES-256 encrypted disks. Live and batch calls can be marked up for redaction, and Scanner allows even raw calls to be automatically stripped of most PII data. And Gridspace Sift makes exporting audio, easy, secure, and completely unlimited.

Glossary of Terms

ASR	Automatic Speech Recognition. Technology that turns an audio recording of speech into a text transcript.
Accuracy	The chance that a machine output conforms to the correct or desired output. Often presented as a "rate" or percentage. 50% accuracy means something is correct half of the time. High accuracy means something is typically correct.
Acoustic	Related to sound. Often used to contract with semantic properties of speech (which are related instead to meaning).
Alert	An instant notification that a result was found in a live speech stream.
Anomaly Detection	A model that informs the user that data is unusual or unlikely. In speech, anomaly detection may find a phrase or full call that's notably unusual.
Call Grading	A set of neural network-based models developed at Gridspace which listen to both the acoustic and semantic content of calls to assign it a "grade". These models are trained from thousands of example calls graded by humans.
Classical Topic Modelling	A term Gridspace uses to distinguish its "extractive" topic feature from what the NLP community often calls "Topic Modelling" (a technique for grouping documents or words into abstract "topics" or clusters).
Classification	A machine learning model which places data into one of several categories.
Compliance	The degree to which an individual or a company adheres to legal, contractual, or company rules. Call center calls have many important rules to track and enforce, often nearly unenforceable in all calls.
Conversational	Two or more people speaking in a natural way to exchange information and achieve goals. This contrasts with high structured speech like someone might use when giving commands to a voice assistant of IVR system.
Corpus / Corpi	A large collection of text data.

Deep Neural Network (DNN)	A type of machine learning algorithm in which a network of mathematical transformations are performed in many layers. These models work well on data that is largely organic and unstructured like pictures, audio, and text. However, DNN's require large datasets to train.
Emotion	In the context of this document, emotion refers to models that detect whether a speaker has a strong emotional coloration in their speech (anger, frustration, joy, relief, etc). This contrasts both to non-emotional acoustic features (dialect, gender, loudness) and semantic content.
False Negative	In a binary model, an output which incorrectly indicates that a particular attribute is absent. An example might include a search query for the word "medical" that misses an instance when the word was spoken.
False Positive	In a binary model, an output that incorrectly indicates that a particular attribute is present. For example, if a model is supposed to detect a call with a low grade, and a call with a high grade is detected.
Keywords	In this document we used keyword detection to refer to systems that can only detect specially coded phrasing. For example, some competing products have keyword detectors for some negative words, but are unable to provide a general transcript.
Latency	The time it takes for an operation (ie, transcription of a word) to occur after the input is generated (ie, a caller says a word).
Live	Nearly instant or with negligible latency. This contrasts with batch or post-call systems that only provide insights at a later time.
Machine Learning	A class of software algorithms that improve over time as they are presented with more data.
NLP	Natural Language Processing is a blanket term for algorithms that extract information and meaning from natural language. Natural language contrasts with formal or machine language, where a person must follow unnatural rules and structure to be understood. A human conversation is natural language. A computer program is not.

Phonetic	Related to how a word sounds.
Phonetic Indexing	A legacy form of audio search where audio is indexed by rough sounds in an audio recording. It allowed for searching audio when high quality transcriptions could not be generated. It can also refer to technologies like metaphone which index text by their phonetic pronunciation. Sift uses these algorithms to aid in searching for words (like names) with many spellings.
Prosodic	Related to how something is said (tone, cadence, intonation).
SIP	Session Initiation Protocol.The standard used for starting and managing telephony calls over IP networks, and, increasingly, for integrating with all modern telephony systems as well.
Scanner	The Gridspace technology that allows for encoding complex queries and rules and searches or alerts. Scanner is flexible to rephrasing and is also sensitive to generic concepts like names or dates, as well as some prosody and sentiment features.
Search Index	A system that is designed to quickly and accurately lookup data in a large dataset. Often the data is organized in advance so searches are fast and flexible.
Semantic	Related to the meaning of words.
Sentiment	The extent to which speech or text is positive, negative, or neutral.
Speech	Audible, meaningful communication generated by a person's speech organs (mouth, larynx, and lungs).
Summarization	Distilling communication into a shorter, more compact form. Intended to give the gist of what was discussed in a conversation.
Supervised Learning	A machine learning algorithm that requires many examples of correct outputs given an input.

Telephony	Speech that comes from cell phones, landlines, or internet calls.
Text	Data that records written communication (like chat, documents, email, or machine transcripts).
Text Analytics	Analysis software that was designed primarily to process text instead of speech (or its transcription).
Topics	The important words and phrases in a conversation.
Training Data	Data that is gathered and prepared so that a machine learning algorithm can use it to improve at a task.
Transcripts	Text that attempts to capture the words spoken by a person or people.
Unsupervised Learning	Machine learning that learns from a large set of data that is not labelled with a correct output.
Word Cloud	An arrangement of words used in a conversation or text, in which the size of each word indicates its frequency or importance. A primitive form of text analytics.
Word Error Rate	The percentage of words in a transcript that are wrong. More formally, the number of insertions, deletions, and substitutions divided by the number of words in the correct transcript.