VOCI

Words, Not Sounds

How a full-text ASR system improves GRC

Executive Summary

- Governance, risk management, and compliance (GRC) will struggle to detect inappropriate behavior in voice conversations that violates standards, policies, and laws. Thus, an Automatic Speech Recognition (ASR) system is a necessity.
- An ASR system processes language just like the human brain: by receiving and encoding an audio input, subjecting it to internal processing, and producing a linguistic output.
- Some ASR systems are phonetics-based, which produce a set of sounds. Other ASR systems are full-text, and produce a set of words and sentences.
- For GRC purposes, a full-text ASR system is preferable, as it is easier to search, accommodates new vocabulary, and affords opportunities for deeper analysis.

INTRODUCTION

Governance, risk management and compliance (GRC) covers a wide range of practices and needs. In general, good GRC practice requires behavioral analysis of employees, contractors, and other individuals for whom a business has responsibility. If this analysis cannot be performed, then it is almost certain that inappropriate behavior that violates standards, policies, or even laws is happening somewhere.

Most business communications (such as email and social media) are easy to analyze automatically for violations. The one exception is audio communications, which are most often phone calls. Even if phone calls are recorded automatically, it requires an inordinate amount of resources to analyze every one made and received by a business. Automatic Speech Recognition (ASR) systems are the obvious solution to this problem, but not all ASR systems have the same benefits for GRC.

LANGUAGE PROCESSING: THE BASICS

To understand which type of ASR system is best suited for GRC purposes, you need to have a basic understanding of how ASR systems process language.

Natural Neural Networks

Natural neural networks, otherwise known as human brains, process audio for speech recognition in three stages.

First, the **input,** a sound wave, is received by the ear. The input is then encoded in a form that can be processed by the brain.

Second, upon receiving the signal, the brain **processes the input** in order to determine which sounds and words are being said. The exact details of this are complex and still a matter of dispute among neurobiologists.

Third, the brain delivers its **output**, an assessment of what is being said, in the form of words or sounds that are recognized by conscious awareness.

Artificial Neural Networks

Artificial neural networks in ASR systems handle audio in a manner similar to natural neural networks, using artificial intelligence for input, processing, and output. The original input (a sound wave) and final output (words or sounds that we can recognize) are identical. However, the encoding of the input into a form that the artificial network can process is quite different, because artificial networks require inputs to be encoded as numeric data that is processed by software, rather than as neural impulses processed by the brain.

Transforming a sound wave to a set of numbers — without losing important information — requires a complex mathematical calculation of two measurement sets. First, the sound wave is sampled, by measuring and recording the height of the wave at regular intervals. Second, the samples are pre-processed to measure the energy contained in each frequency band. This results in a set of numbers that is difficult for us to recognize as sound without their visualization. It is, however, ideal for an artificial neural network to begin its work.

Conceptually, the neural network takes the encoded audio input and compares it to a dictionary of known sounds or letters to determine what is being uttered at each sampled point. After each point is assessed, the network uses that information as additional input to the assessment of the next sampled point, reflecting the fact that certain sound and letter combinations are more likely than others in human languages. Thus, the neural network is **recurrent.**

Once a set of sounds or letters is produced, a final refinement is performed, which eliminates duplicate letters and sounds that are unlikely or incorrect in the language being processed (e.g., "JJ" or "SSS" in English), as well as removing any blank areas where the network was unable to recognize any sound or letter in the input.

The output stage is the same as for a natural neural network. It is an assessment of what is being said, in the form of words or sounds that can be recognized by humans.

PHONETICS-BASED AND FULL-TEXT ASR SYSTEMS

ASR systems are either phonetics-based or full-text. The distinction between the two systems resides in the processing stage of their artificial neural networks, which determines the type of output.

A **phonetics-based ASR system** does not process encoded audio inputs into a set of letters at all. Instead, this system uses a set of **phonemes**. A phoneme is the smallest perceptually distinct sound that exists in a language, like the "kuh" sound at the end of "work". Therefore, a phonetics-based system's output is a set of sounds.

A **full-text ASR system** uses phonemes as an intermediate step, but then compares them to a dictionary of words and phrases, using a **large vocabulary continuous speech recognition (LVCSR) engine**. The output of the system is a set of words or sentences from the dictionary.

BENEFITS OF FULL-TEXT ASR SYSTEMS

There are many more words in a language than there are phonemes. Consequently, the dictionaries of full-text ASR systems are much larger than those of phonetics-based ASR systems. This means phonetics-based ASR systems have a processing speed advantage over full-text ASR systems. However, this is where their advantage ends. Full-text ASR systems provide for accommodation of new vocabulary, easier and more flexible search capabilities, greater transcription accuracy, and in-depth analysis, making them superior for GRC purposes.

Phonetics-based ASR systems can learn new sounds, but they can never learn new terms like full-text ASR systems do. They will output the way the terms were pronounced without recognizing the terms' significance, thus producing "transcripts" that are not as useful. Even if such "transcripts" are fed to a separate analytics tool, there is no way to accurately determine which phonemes belong together as a word. If no words can be identified, new or unusual words cannot be identified, either.

In comparison, unusual or non-standard names, initialisms, acronyms, and industry-specific jargon can be added to the dictionaries of full-text ASR systems to ensure that they are correctly recognized and transcribed. This language extensibility is especially useful if an organization needs to identify emerging trends in conversations, such as terms or phrases that were not considered in advance or words that are unusual.

It is remarkably difficult to search phonetic outputs in phonetics-based ASR systems if a word can be pronounced differently ("laboratory") and if different words can be pronounced the same way ("there", "they're", "their"). Any search would produce a number of false positives and false negatives, requiring a significant manual

review to ensure that every result returned is actually relevant. Full-text ASR systems, on the other hand, perform an algorithmic assessment of which sound-alike words are most probable in context, creating transcripts with much fewer errors.

In-depth analysis can also be performed on transcripts produced by full-text ASR systems. An analytics tool could take a transcript of a recorded conversation and generate a word or sentence cloud of the most frequently spoken terms or phrases, or identify words or sentences that are being used with increasing (or decreasing) frequency across a certain period of time (e.g., the past month). What's more, sentiment analysis can be accomplished with full-text ASR transcripts. Like a phonetics-based ASR transcript, a full-text ASR transcript can be marked by the transcription system or a separate analytics tool to show changes in tone, pitch, and volume. However, a full-text ASR transcript also indicates more subtle changes, such as when a caller is becoming increasingly frustrated or angry. The caller may not raise his or her voice, but will use words and phrases to convey negative sentiment. Full-text ASR systems can create a transcript suitable for a more thorough analysis, and make it possible to recognize words and phrases that have negative connotations, even without clear acoustic markers.

APPLICATIONS FOR GRC

GRC practitioners require a robust ASR system that enables them to search for key terms and phrases that are not only relevant during the time of deployment, but also when responding to future issues. In addition, detailed analysis of the system's transcripts should identify potential risks to an organization and possible compliance violations, and help facilitate and expedite the implementation of corrective measures. A full-text ASR system meets all these needs.

As noted above, a key terms search in a phonetics-based ASR system is limited to those terms which can be anticipated in advance and rendered into phonetic form. A financial institution, as an example, may want to watch for the word "credit", and an ASR system could recognize the specific phonemes (approximately, "kuh-reh-di-tuh"). This may be successful if one can be sure that only a handful of terms and phrases is relevant.

However, most organizations will need to identify terms or phrases that were not considered in advance, and words that are significant but uncommon, such as "risk" or "illegal" or "sue". Generally speaking, while every organization can identify some GRC-related words in advance, there is always a likelihood that new issues will emerge that must be acted upon quickly to protect the organization.

Additionally, a full-text ASR transcript provides GRC practitioners with more usable data for analysis, whether performed manually or by a separate analytics tool. Word clouds and trending vocabulary can be identified, which enable a clearer understanding of potential legal and compliance ramifications in caller/agent conversations. This means that risks to the organization can be identified with much greater speed and accuracy.

Similarly, with a deeper sentiment analysis, a clearer identification of potential risk exposure and potential compliance issues can be made. For example, an angry caller may be more likely to take a consumer or legal action against an organization, or an agent who is increasingly using negative towards customers may be in violation of company policies.

CONCLUSION

Artificial neural networks process language in much the same way as our brains do. However, there are key distinctions in how the audio input is encoded and in how the input is processed by the network.

From the perspective of GRC, full-text ASR systems offer greater language understanding, search capabilities, transcription accuracy, and analysis than phonetics-based ASR systems. In addition, full-text ASR systems make detecting GRC violations in voice not only possible, but also efficient and effective.

voci

(412) 621-9310 | info@vocitec.com | www.vocitec.com | @Vocitec