White Paper

Information Management & Governance

Maximize Your Audio Assets with Next-Gen Technology: An Overview of Speech Technology

MICRO FOCUS

Table of Contents

The speech analytics market is estimated to reach $1.33 billion in 2019, at a compound annual growth rate (CAGR) of 23.9% from 2014 to 2019.

marketsandmarkets.com

The Rising Importance of Understanding Speech

The amount of audio data is growing every day, creating an increasing need to be able to search and analyze it. Whether the data comes in the form of call center recordings, voice-activated search commands like Siri, or audio in video assets such as Web conferencing or broadcast streams, having accurate speech analytics technology on board has never been more critical.

Traditional approaches, however, are not designed for the Big Data era. Simply tagging rich media with key concepts or relying on metadata gives inconsistent, incomplete results. These approaches are also manually intensive and not scalable. The explosion of audio assets requires real-time, automated, integrated analysis similar to that available for other forms of data. Niche solutions that provide audio analysis typically handle the function separately from non-audio content, making it impossible to get a holistic understanding of your data. Micro Focus® IDOL Speech provides organizations with an automated, scalable solution for understanding audio and speech data, without the need for time-consuming manual processes. We call this augmented intelligence, which makes use of many artificial intelligence techniques.

Artificial Intelligence as an Aid to Decision Making

The realization that computers could perform calculations that would take humans many weeks, or that we could not achieve at all, rapidly engendered the idea that we could make these mere machines do all we can do. It is no coincidence that the golden era of science fiction began around the time of the great burst of computing and space technology that followed the Second World War. In much of it, the futuristic visions include nonhuman technologies whose capabilities far outstretch those of their mortal creators.

Artificial intelligence is already all around us. The sensors that help determine the optimal time to change traffic lights, washing machines that automatically adapt to the quantity of clothing, and the ever-changing gameplay of our favorite smartphone games are examples. Even the system that prevents a microwave from starting because the door is open is artificial intelligence in action. Computers are good at making decisions when armed with all the relevant information and with no element of randomness—systems known as "fully observable deterministic systems." For this reason, computers are far better than humans at playing checkers but fall down in the very human game of poker.

But while we strive for a world in which a machine will keep our garden perfectly manicured or tell us exactly when to book our vacation to guarantee perfect weather, most observers of the field of artificial intelligence fail to note one crucial aspect: When it comes to the most human parts of our existence—our interactions, decisions, and interests—machines have far less to offer. We don't actually want the machine to pick which flowers to plant, we just want it to execute our wishes quickly and efficiently. Less still do we want a computer to tell us **where** we will be taking our vacation, even if we are clamoring for it to help us make a smart and informed choice.

From its inception, IDOL has created a number of pioneering techniques in artificial intelligence that help to automate and enhance the processing of human information of all types. Our aim is not to take the decision away from humans, but to equip them with the information to determine the best course of action. We call this approach to artificial intelligence augmented intelligence.

Within augmented intelligence, IDOL uses a wide variety of theories and techniques to process and extract meaning from human information. Let's look at how some of these theories and techniques are applied to speech analytics.

IDOL Enables Accurate Speech Analytics Using Deep Learning Technology

IDOL Speech, powered by the IDOL Speech Server, makes rich media content searchable, the same as any other text-based content. This unique concept-based speech analysis forms an understanding of the content as easily as for text-based content. At the forefront of the speech recognition research community, IDOL technology integrates deep learning, through the use of artificial neural networks, to enable highly accurate speech operations. Advanced speech technology is just one aspect of what the IDOL platform offers to enable you to unlock valuable insight from your content. IDOL also encompasses other rich capabilities such as image, video, text, and structured data analysis to help you harness data of virtually any format. While converting speech content to text creates immense value on its own, you can gain additional insights when using it with other IDOL operations such as content categorization, entity extraction, and sentiment analysis.

Monitor, Search, Access, Analyze, and Understand Audio

IDOL Speech allows users to monitor, search, access, analyze, and understand audio and video data from virtually any source. IDOL Speech provides a deep index for speech and audio content. A typical activity, where IDOL processes unstructured data, consists of collating and ingesting content, enriching and augmenting that content, performing operations and analyzing the content, and then publishing the data for your audience.

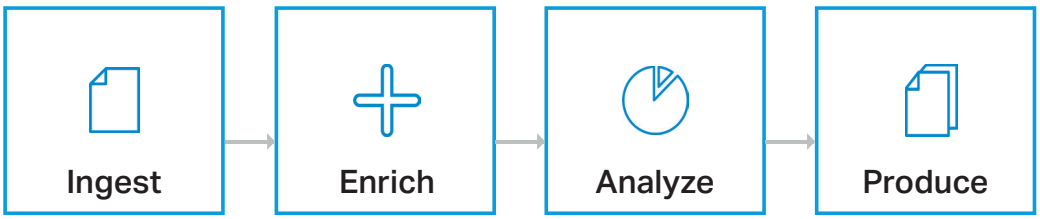

Figure 1. Key processes of IDOL Speech Technology

IDOL Speech has been used for analyzing speech for over 15 years and has been deployed across numerous sites for a variety of uses. The ability to customize the language to a specific domain has led the way for analytics on tough audio such as voice calls. Now, with the adoption of next-generation algorithms, IDOL Speech enables customers to apply speech analysis capabilities over a wide range of applications—out of the box and without any customization.

IDOL Speech is part of the IDOL platform and delivered as an ACI server. IDOL Speech encompasses several speech and audio analysis operations, including those listed in the following table:

Type	Function	Description
Speech	Speech to text	Converts spoken speech to a text transcript of the most likely words
Speech	Phonetic phrase search	Converts spoken speech to a phonetic index, which can be searched
Speech	Speaker segmentation and identification	Identifies the speakers in spoken speech
Speech	Spoken language identification	Identifies the language spoken
Speech	Transcript alignment	Aligns a given text transcript with an audio file producing time stamps for all words
Audio	Audio quality and classification	Classifies audio segments as music, noise, or speech, and gives details on the audio quality
Audio	Audio fingerprint identification	Allows creation of an audio database for identifying audio segments
Audio	Audio security	Identifies common security threats from audio captured
Model	Language customization	Allows customization of language models used in speech-to-text operations
Model	Acoustic adaptation	Allows adaptation of acoustic models used in speech-to-text and phonetic phrase search operations

Table 1. Major functions and descriptions

Advantages of IDOL Speech

- All speech operations are asynchronous, allowing instant access to results from audio and speech operations. IDOL Speech can process both audio from files or over a stream, allowing real-time reporting so events can be flagged immediately, meaning you don't have to wait for the end of a file or stream event.
- IDOL Speech supports multiple languages. A single instance of IDOL Speech can process several languages simultaneously.
- IDOL Speech lets you put together combinations of speech processing functions to create custom operations, allowing you to perform several processes simultaneously on audio data.

Combining IDOL Operations

With IDOL Speech, you can convert audio and speech into textual information that you can then combine with other IDOL operations, such as concept extraction or sentiment analysis, to gain further insight and analysis of your audio content. You can combine and chain IDOL operations to produce workflows that enable you to really pinpoint key audio content that enables you to make informed decisions for your organization.

Over a large corpus of unknown spoken audio, you can determine the most likely language spoken, discard areas of music and noise, and then build a searchable archive using IDOL Speech to Text and indexing to IDOL Server.

You can then augment your analysis with further IDOL operations, including:

- Categorize your rich media content according to topic e.g., transport, sport, weather news
- Discover overtly positive or negative sentiment over calls in the call center
- Pull out company or product names as well as addresses and customer telephone numbers using entity extraction
- Automatically cluster your rich media content by topic to identify the top themes in the data

The Micro Focus Broadcast Monitoring solution uses IDOL speech and video analytics technologies, along with other IDOL operations, to provide a deep understanding of broadcast video content, making it possible to perform operations such as search and retrieval over a large corpus of broadcast content.

At the core of IDOL operations is its conceptual analysis.

Conceptual Analysis

The probabilistic approach of IDOL index and retrieval process allows complex operations to occur naturally. This augments basic retrieval to allow more subtle connections to be determined and more relevant results returned than are possible in any keyword engine.

As an example, imagine we are interested in the effect of pollution on penguins. The traditional approach to finding information to satisfy our interest would be to select a keyword search engine and type in the word **penguin**. This would return useful content, but also a significant amount of irrelevant content about a publishing company, the chocolate biscuit, Batman and Robin, and so on.

In our case, however, we are interested in content that has a high probability of being about penguins, the birds. A document containing the word sea could be about penguins but sea occurs in many contexts, and therefore there is a significant probability that the content is about something else. However, if a document

contains the words black, white, flightless, feather, slick, and oil, the probability that the document is not about penguins and pollution becomes extremely low. Furthermore, this has been identified without using the word penguin and instead using a larger amount of weaker information, any of which can be taken away without significantly affecting the probability. The Micro Focus approach understands context based on either strong concepts and keywords or a larger amount of weaker information.

To do this, IDOL needs a framework to encapsulate concepts such as penguins, the birds, or news about the weather. For this purpose conceptual agents are used.

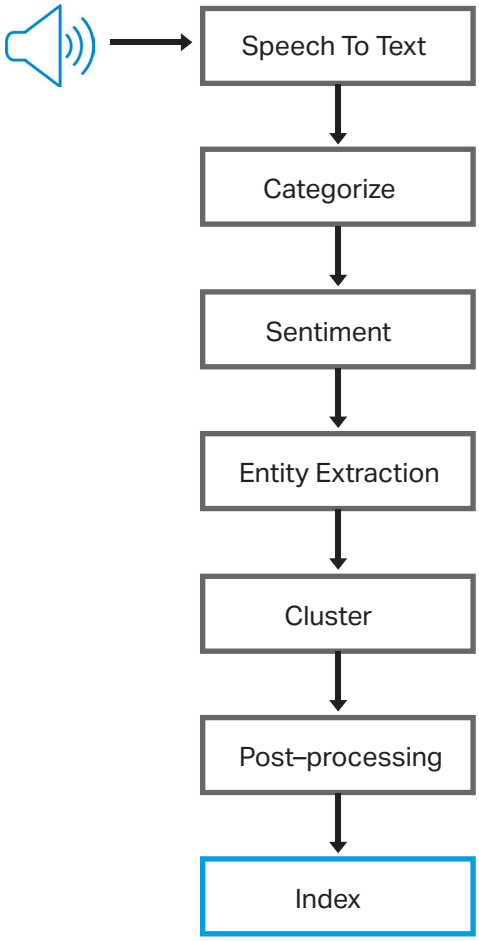

Figure 2. Broadcast monitoring in action

Speech to Text Using Deep Learning

Speech to text is the process of translating spoken words into text and includes a number of ways to analyze, search, and process audio content, such as:

- The command and control of mobile devices
- Interactive voice systems for automated call handling
- Dictation of letters, memoranda, and other electronic text documents
- Audio and video search, where the search for specific terms or concepts is performed on the transcript
- Subtitles or closed captions for video
- High-level analysis of phone calls in contact centers

IDOL Speech delivers state-of-the-art speech to text using deep learning through the use of artificial neural networks, which deliver far more accuracy than statistical algorithms. It is trained on many hours of sample speech and language data to "learn" the patterns of speech. This training process produces language models, which make up the IDOL Speech language packs. Both the acoustics and linguistics of a particular language are modeled; for instance, the acoustic model finds probable (phonetic) speech sounds in the speech audio, which is then combined with the lexicon and language model to find the most likely sequence of words and phrases. The acoustic model uses deep neural networks (DNNs) to perform deep learning of speech.

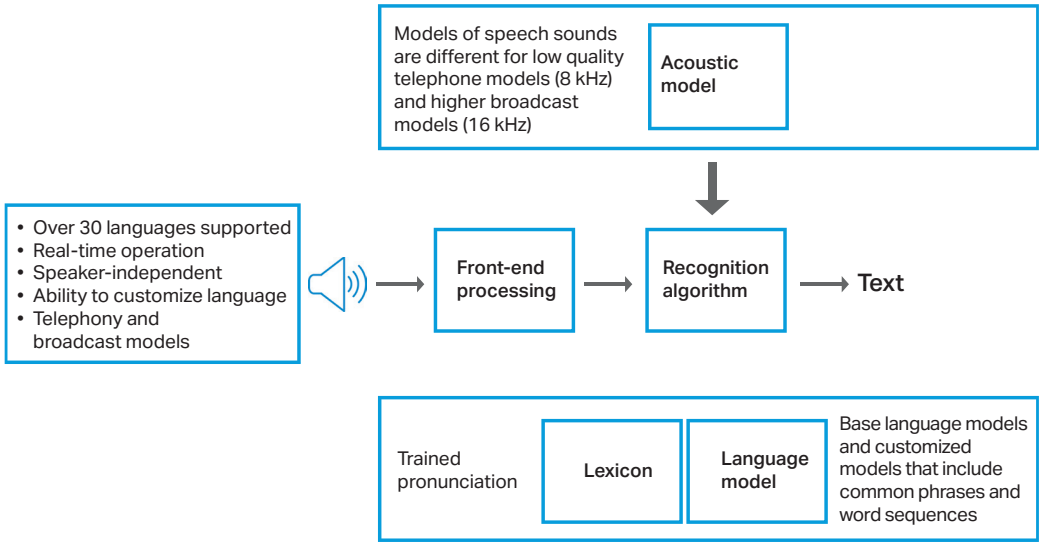

Figure 3. Speech-to-text architecture

Language Customization and Acoustic Adaptation

IDOL Speech requires language packs to perform speech processing tasks. A language pack contains a language model and an acoustic model. The key components of the language model are:

- The word vocabulary and the pronunciation dictionary
- The word n-gram probabilities

The IDOL Speech language model covers a broad vocabulary, reflecting the general spoken language. However, in some cases, the need arises to process speech data covering specialized topics, such as financial or medical topics, or atypical sentence structures. In these cases, custom language models can be built for IDOL Speech to use when processing this audio. Building a new language model requires a lot of text—in the order of millions or billions of words. The standard language packs are usually built with many billions of words of text. Therefore, the best way to customize a language model is to build a small custom language model that uses the specialized text, and then combine it with the standard language model when you perform speech to text.

In addition, IDOL Speech allows you to adapt acoustic models available out of the box so that they more closely match the acoustic properties of a particular audio data set. Adapting the model using data that closely represents the audio you are processing should improve speech-to-text results in terms of recording quality and accents. This adaptation technology is only available for our statistical-based acoustic models, due to special-purpose hardware required for deep neural network training. However, the need for adaptation is diminished due to the increasing adoption of next-generation algorithms.

We are continually updating our language models as more training data becomes available. Often, customers provide materials to help improve speech analytics performance, via tools that preserve the confidentiality of the audio and transcript.

Speaker Segmentation and Identification

Speaker segmentation and identification is the process of segmenting speech by speaker identifying speakers based on their unique voice characteristics.

You can train the system to identify certain speakers using samples of speech from each speaker to create speaker templates.

IDOL Speech also comes with default templates that identify speakers by their gender alone. This means that even without any training, you can see speakers in the spoken audio and their gender.

IDOL Speech also provides operations to merge speaker segments found by common voice traits, allowing speakers to be found automatically without the need for training.

You can also use speech clustering to segment a speech waveform and separate it out into a number of speakers. IDOL Speech produces a timed sequence of labels that correspond to speaker assignments.

Spoken Language Identification

Spoken language identification is the process of determining which natural language is being spoken, but does not require the identification of the spoken words in the content. To perform spoken language identification, IDOL Speech first distinguishes the speech sounds, phonemes in speech, and then chooses a language that has the closest distribution of phonemes. Spoken language identification is text-independent.

IDOL Speech is able to detect a number of languages out of the box. You can further expand IDOL Speech language identification by building your own language classifiers. These are trained with speech samples in the relevant language.

Transcript Alignment

Transcript alignment assigns time codes to all words in an audio transcript file. The transcript alignment function processes most transcripts, even if they contain noise and missing sections. The generated time codes are normally accurate to within half a second. Transcript alignment is useful for:

- Generating subtitles for videos from manual transcripts
- Creating time indexes for words in the transcript so that the audio can be searched and positioned

Speech to text is also used in the process of generating the time codes.

Audio Quality and Classification

IDOL Speech allows you to automatically classify audio as music, noise, or speech. This can be useful when you are running the speech-to-text operation on audio content that may contain music. You can combine these operations such that a speech-to-text transcript is only produced for those audio segments classified as “speech.”

In addition, IDOL Speech performs a number of other operations on audio content such as computing the signal-to-noise ratio, identifying clipping of the audio signal, and many others, to determine the quality of the audio.

Audio Fingerprint Identification

Also known as acoustic fingerprinting, audio fingerprint identification generates a digital summary of an audio sample, to identify it quickly or to locate similar samples in a database. You can use audio fingerprinting to:

- Identify songs, melodies, and jingles
- Identify advertisements
- Identify media tracks, where the media track can consist of human voices, music, or any other discernible sounds; for example, searching for the phrase "President Obama's inaugural speech."

The audio sample to be identified does not need to be an exact copy of the original.

Audio Security

Audio security detects and labels segments of audio that contain security-related sounds, including:

- Various alarms, including car alarms
- Breaking glass
- Screams
- Gunshots

Phonetic Searching

Phonetic search is the process of searching for words and phrases by their pronunciation, and the use of phonemes. Phonemes are the fundamental units of sound that make up a spoken language. For example, the word catch contains three phonemes, or sounds, k–a–tch.

To perform phonetic search, IDOL Speech phoneme identification engine first processes an audio file to create a time track of phonemes, which reports the time each phoneme occurs in the file. This is a one-time process. IDOL Speech then searches the phoneme time track data for the specified words or phrases. On an average desktop computer, the search process can operate a few hundred times faster than real time.

The best method for performing a phonetic phrase search is to perform a full speech-to-text operation, which opens up the full set of IDOL analytics operations, including conceptual search, for further analysis—versus just a phonetic phrase search. However, there are cases where you may have specific requirements to use keyword and phrase identification and want to limit hardware resources. Phonetic phrase search can be used in those instances. Also note, phonetic phrase search is language-dependent.

Comparing Phonetic Search and Speech to Text

Speech to text offers a more effective approach than phonetic search in almost all cases because it uses the same acoustic models as phonetic search. It also adds language modeling to provide a higher level of lexical, grammatical, and semantic intelligence. These capabilities steer the recognition to better performance than purely acoustic phonetic search. For instance, this method can determine that "Barack" is highly likely to be followed by "Obama," and preceded by "President."

The advantage of phonetic search is that it allows a trade-off between false positives and false negatives. For instance, you can make the phrase "Barack Obama" occur in many places on the off-chance that you won't miss the true occurrence, even though there may be 10 to 100 times as many false occurrences than you need to plow through. For this reason, the phonetic search functionality within IDOL Speech is recommended for cases where the audio is poor and the user does not mind sorting through the false positives. So, it doesn't recognize more accurately, but it recognizes more instances of the target whether they're right or wrong. With the recent improvements in speech to text technologies, the case for phonetic search diminishes when there are fewer cases where recognition is sufficiently poor. If false positives and false negatives are equally bad, speech to text is the better solution for good audio and for poor quality audio alike.

Phonetic search can be slightly quicker at index time, though it takes longer at search time as more work remains to be done. Another potential advantage of phonetic search is its open vocabulary, which allows you to search for a word that was unknown to you at index time. While conceptually useful, there are not many instances in practice where the word is either not in the default language models that are trained on billions of words, or you don't know the words at index time. Also, content can always be re-indexed, if required. Similarly, it also means the results of indexing a file for phrase search can be a rather large phone lattice where speech to text indexes into plain text.

Speech to text indexed into the IDOL platform IDOL Server provides another level of intelligence. However, to cover the range of conceptual matches that IDOL Server provides on the speech-to-text output with a phonetic search is simply not possible. Any other technology approaching this level of capability would require a large number of word and phrase combinations, and specialist knowledge of the concepts within the content.

Standalone Server Architecture

IDOL Speech is a standalone server that uses the Autonomy Content Infrastructure (ACI) Client API to communicate with custom applications. It allows data to be retrieved over HTTP using XML and can adhere to SOAP (Simple Object Access Protocol). It supports both synchronous and asynchronous actions (see the section titled Scalability and high performance).

The following operating systems are supported:

- Windows (XP and later)
- UNIX (RHEL, SUSE)

Scalability and High Performance

To improve performance in a production environment, IDOL Speech supports both synchronous and asynchronous actions, and can be distributed for horizontal scaling.

With a synchronous action, IDOL Speech will run the task immediately and return a result when the action is complete. Asynchronous actions allow a user to send multiple tasks all at once, returning a task ID/token for each job. IDOL Speech will then queue the tasks and run them in order. The user can check on the progress of each task, kicking off additional tasks if necessary, to enable better batch processing and more complicated workflows.

In large systems where a very large number of documents need processing, it is possible to distribute work among multiple instances of IDOL Speech using a distributed action handler.

Appendix: Speech-to-Text Languages

The official supported language for IDOL Speech to text:

- English
  - American
  - British
  - Australian
  - Canadian
  - Singaporean
  - Irish
- French
  - European
  - Canadian
- German
- Spanish
  - European
  - North American
  - South American

- Italian
- Danish
- Dutch
- Swedish
- Flemish
- Portuguese
  - Brazilian
  - Portugal
- Welsh
- Catalan
- Polish
- Greek
- Romanian
- Czech
- Slovak
- Hungarian
- Russian
- Japanese
- Korean
- Mandarin
- Hebrew
- Arabic
  - Modern Standard
  - Gulf
- Farsi
- Urdu

Learn More At
www.microfocus.com/idol
www.microfocus.com/richmedia

Additional contact information and office locations:
www.microfocus.com

www.microfocus.com