verbit

Speech Dictionary & Related Terms

Helpful speech-focused lingo and terminology to know

If the terms AI, ASR and NLP have you frazzled, you're not alone. More speech-related acronyms and technologies seem to be popping up regularly, and they're making their way into the daily vernacular.

Here's a helpful cheat sheet of terms for you to reference. Utilize this list to gain a better understanding of common phrases being used as speech-to-text lingo enters your world and workplace.

Acoustic Model:

This model refers to mapping audio signals and the phonemes or other linguistic units that make up speech. The model is trained using audio recordings and their related transcripts in text format. These are used to create statistical representations of sounds that make up each word.

Accuracy:

In this realm, accuracy refers to the correct amount of predictions made by a various speech model or human assisting. Greater accuracy translates to stronger performance. Individuals who rely on speech to text tools, such as those with disabilities, need high levels of accuracy (99% according to the ADA) to comprehend information well and receive equity.

Americans with Disabilities Act (ADA):

The Americans with Disabilities Act works to safeguard the rights of people with disabilities. Businesses with 15 or more employees, universities and cultural institutions based in the US can face legal consequences if they do not adhere to ADA guidelines. Title I of the ADA requires that employers refrain from discriminating against and offer accommodations for individuals who are disabled or potential employees. Here are tips to meet ADA guidelines.

Artificial Intelligence (AI):

A wide-ranging branch of computer science which involves building and using smart machines that are capable of performing tasks that typically require human intelligence. The idea stems around training computers or robots to adopt intellectual processes which are characteristic of humans, including the ability to reason, sense meaning, generalize or learn and correct from past experiences. For example, AI is being used to transform audio that's heard into written words. With greater use, AI can mature to perform more intelligently.

Audio Description (AD)

Audio Description is designed to offer blind and low vision individuals equal opportunities to consume video or visual content. This service is typically powered by professionally trained describers who analyze the visual elements of videos to deliver clear and concise descriptions of visual images as seen by individuals with sight. AD can be offered with either human or synthetic voice description and via an accessible Smart Player in Verbit's case. **Standard audio description** adds the spoken description of visual elements to the natural pauses of video. **Extended audio description** temporarily pauses the audio and video to allow information to be delivered when pauses in dialogue are insufficient for adequate description.

Automated Captioning & Transcription:

Also known as automatic captioning or automatic transcription, this refers to using voice and speech recognition technology to produce captions or transcripts. Automated transcription can be applied on both live and recorded content to provide a visual representation of the dialogue heard. You can think of it as the built-in closed captioning (CC) button on Zoom or YouTube. While automatic sounds ideal and are produced quickly in real-time, the accuracy levels typically are less than ideal. Automatic technologies alone often cannot detect speech in the same way humans can, so they require review and editing. Automated solutions alone typically do not offer an equitable experience, but they're absolutely better to have than not to make content more accessible. If accuracy is your goal, they can be helpful in providing initial captions and transcripts that you can edit manually after.

Automatic Speech Recognition (ASR):

Technology that transforms speech, or an audio signal, into text. It uses knowledge of linguistics, computer science and electrical engineering to produce the text. It's often used as the basis for automatic captioning and transcription solutions. Again, for high accuracy levels or to meet ADA guidelines, users typically need to edit the technology's work after or use a system like Verbit's which uses ASR + human intelligence through editing to provide highly accurate results.

Captioning:

Captioning refers to the process of converting audio content into text and displaying it on screen or on a nearby monitor. It can be applied to recorded videos, live TV broadcasts, virtual event live streams, films, podcasts and more. Captioning enables those who are Deaf or hard of hearing to have access to media materials that they likely could not consume otherwise. While captioning was initially designed to serve viewers with hearing loss, everyone can benefit from the additional visual aid to capture and retain what was said. Many individuals without hearing loss prefer to watch their content with the captions on. It's also helpful for individuals who cannot watch with sound, non-native speakers watching content in other languages and those with ADD and ADHD to help them maintain focus.

Closed Captions:

Closed Captions (CC) allow viewers to select whether or not they want to watch with captions. They're initially hidden and are made visible by a decoder at the time of viewing via a remote control or an on-screen menu which features a CC button. They are typically displayed in white capital letters encased in a black box.

Deep Learning:

Deep learning is a machine learning method based on artificial neural networks. Deep learning is designed to simulate the way the human brain works. It can be supervised, semi-supervised or unsupervised. As deep learning aims to mimic the human brain it enables systems to group together data and make predictions with remarkable accuracy.

Deep Neural Network (DNN):

A DNN is a neural network with more than two layers, allowing for a certain level of complexity. Deep neural networks offer great statistical value, increasing the accuracy of a machine learning model. Deep neural networks use complex mathematical modeling to process data. The DNN finds the correct mathematical manipulation to turn the input into the output. The relationship can be linear or nonlinear.

DEI:

Diversity, Equity and Inclusion are often now grouped together and abbreviated as DEI. DEI refers to efforts being made by brands, businesses, universities and cultural institutions to drive and promote inclusion. In the speech world, captions, transcripts, audio description and translation are all seen as helpful tools to promote more accessible and inclusive cultures and are being implemented as part of DEI policies.

Dubbing:

Dubbing is the process of creating a new spoken audio track for a video in a different language and adding it to the original video as a replacement for the original speaker's voice. Dubbing allows viewers to follow the audio in their native tongue. The best dubbing utilizes talented voice actors who can capture the emotions, tone and idiomatic expressions of the original audio recording. Dubbing should be produced and timed to fit seamlessly into an existing video.

FCC:

Being FCC compliant often comes up with regard to video captioning and speech-related services. It means following the rules and regulations which The Federal Communications Commission (FCC) has laid down for video programming distributors (VPDs), including cable operators, broadcasters, satellite distributors and multi-channel video programming distributors. Congress requires them to not only close caption their programs, but provide a certain level of quality to ensure that viewers with hearing loss, among others, have full access. Per the FCC, captions are required to be: **accurate** (match the spoken words in the dialogue and convey background noises and other sounds), **synchronous** (coincide with their corresponding spoken words and sounds to the greatest extent possible and displayed on the screen at an acceptable for reading), **complete** (run during the entirety of the program) and **properly placed** (not blocking other important visual content on the screen, overlap or run off the edge of the screen).

GDPR:

In contrast to the US, the EU has a comprehensive law related to data privacy. The EU treats data privacy as a human right, and the penalty for violating the General Data Protection Regulation (GDPR) is severe. Non-compliance can result in fees of up to 20 million euros, or 4% of the "total worldwide annual turnover," whichever is larger. The GDPR impacts all companies that collect data from persons in the EU. The reach of the law is broad and applies to foreign companies that do business in the EU. GDPR compliance comes into play when working with vendors or partners and protecting data while using technologies and services to support accessibility measures. For example, Verbit prides itself in being GDPR compliant to give its partners peace of mind when building more inclusive content and aiming to reach wider audiences with its captioning, transcription and translation solutions.

Glossary:

To perform with greater accuracy, some speech-to-text solutions provide users with the option to upload content or key terms into their system beforehand. This glossary can include a list of speaker names with the correct spellings, content from prior videos or events, relevant materials and more complex terminologies that are likely to come up. Having this glossary uploaded into the system or in front of the human captioner prior to the event, allows the terms to be easily detected and spelled correctly, making for the best viewing experience and most professional captioning and transcription result possible in real-time.

GSA:

This term comes up when government bodies and entities are looking to enlist captioning or transcription partners to make their experiences, content and events more accessible. The GSA Schedule, also known as Federal Supply Schedule, and Multiple Award Schedule (MAS), is a long-term governmentwide contract with commercial companies that provides access to products and services at fair and reasonable prices to the government. Government agencies can greatly benefit when working with **GSA Schedule contractors**, such Automatic Sync Technologies, a Verbit company, to fill repetitive needs for captioning and transcription at budget.

HEERF:

The Higher Education Emergency Relief Fund (HEERF) is authorized by the Coronavirus Response and Relief Supplemental Appropriations Act, 2021 (CRRSAA), Public Law 116-260. HEERF Funds originate from the Federal Coronavirus Aid, Relief and Economic Security (CARES) Act of March 2020. They are emergency financial aid grants that have been distributed to assist universities and students across the US since the start of the pandemic. Often, these funds can be applied to speech-to-text solutions, such as captioning and transcription, that assist with accessibility needs for learners with disabilities.

Human Transcription:

Human transcription is conducted by real people who listen to an audio file and convert it to text. Humans often produce more accurate results than AI-based transcription alone, as they can decipher different accents and industry jargon. Human transcribers can also be assigned to work on transcripts produced initially by automatic speech recognition to edit the work to reach the 99% accuracy levels the ADA requires. Verbit employs a team of 35,000 human transcribers.

Language Model:

Language modeling is used in speech recognition by using statistics and probability to determine word sequences. The language model provides context to decipher words and phrases that may sound similar phonetically, but are different. It is designed to estimate the likelihood of different phrases and is useful in reaching higher accuracy levels when producing captions and transcripts.

Machine Generated Transcription:

Machine generated transcription relies on software or ASR to assist in the conversion of human speech into a text transcript. Its performance can depend on the quality of the recording and be generated automatically on both live and recorded video. Typically, it does not perform at the same level of accuracy as human transcription and needs to be edited manually in some instances.

Natural Language Processing (NLP):

Natural language processing refers to a computer program's capability to understand human language as it is spoken and written. It's based on the idea of building machines that comprehend and can respond to text or voice data with text or speech of their own, similarly to how humans do.

Natural Language Understanding (NLU):

Natural language understanding is a part of artificial intelligence that uses software to understand input in the form of sentences using text or speech. It allows for human-computer interaction, allowing computers to speak back to humans in their own language. NLU is used to create chat and voice-enabled bots that can respond to assist humans without supervision.

Open Captions:

Open captions are often referred to as burned-in, baked-on, or hard-coded captions. These captions are seen by everyone who consumes a video they are placed on. Open captions are a permanent feature on the video, as they cannot be turned on and off.

RTMP:

Real-Time Messaging Protocol, or RTMP, can benefit broadcast companies who are streaming events in real-time. It supports what's known as low-latency streaming. It is also well known for minimal buffering, which enhances the user experience. RTMP is a live streaming protocol. Essentially, it's part of the technology that makes live streaming possible. Hundreds of platforms support RTMP, including YouTube, Vimeo, Livestream, Twitch, DaCast, Periscope and Facebook Live.

Speech-to-Text (STT):

Speech-to-text, which is often called speech recognition or automatic speech recognition (ASR), refers to the process of using technology to take spoken language and turn it into text. It benefits individuals who are Deaf or hard of hearing, as well as many others. Many speech recognition systems can be trained to perform at stronger accuracy levels. Speech-to-text also helps with searchability, allowing users to search and find elements from a video or meeting quickly and easily.

SRT:

An SRT file, or SubRip Subtitle file, is a plain-text file with important information regarding captions or subtitles. SRT files include the start and end timecodes of the text to ensure the captions or subtitles match the audio. An SRT file is simply a text file that can be uploaded on YouTube, LinkedIn, Facebook and the like, in addition to the video or audio file it matches.

Subtitles:

Subtitles are text pulled from either a transcript or screenplay of the dialogue or commentary in videos, films, television shows, video games and more that are always displayed at the bottom of the screen (or at the top of the screen if other text exists and will be blocked). They translate or transcribe the dialogue or narrative and are typically always “on” or visible.

Transcription:

Transcription involves the process of listening to audio, video or live speech and writing it out into text format in the same language. Word-for-word transcription captures every word said in the dialogue. Producing transcripts of videos, calls, meetings, events, lectures, training sessions and more can be helpful tools for referencing and note taking. Transcripts are also essential for legal and medical purposes as they provide an admissible, word-for-word record of everything said in a deposition or doctor’s visit, for example. Transcripts can be produced live to appear in real-time within platforms like Zoom and serve as effective notes. Transcripts can also accompany audio or video files to make them searchable and improve SEO.

Translation:

Translation refers to converting text files from one language to another language, such as English to Spanish. Translations can be produced from transcripts of an audio or video in the original language whereby the text is written in full and converted into the desired language. Translation is helpful for companies aiming to reach larger audiences with their content.

Post-Production Captioning:

Post-production captioning is the process of adding captions to a video after it is recorded. These captions must reach 99% accuracy levels to meet ADA guidelines. They should be synchronized to the video, readable and positioned on the screen in areas where they do not cover important information. Post-production captions should be added to all video content in educational and professional environments, as well as on social videos to make them more accessible and inclusive to adhere to legal standards.

FedRAMP:

The Federal Risk and Authorization Management Program is a US federal government-wide program. It provides a standardized take on security assessment, authorization and monitoring for cloud products and services. For example, government agencies aren't using Zoom Commercial, but rather Zoom for Government for their meetings, events and town halls. The platform was granted authority to operate by the Department of Homeland Security and the FedRAMP PMO. Many automatic or speech-to-text solutions which integrate into Zoom Commercial and enhance the experience for participants aren't granted integrations into ZoomGov as they aren't secure enough to meet FedRAMP requirements. It's critical for government agencies to provide captions and transcripts to meet accessibility needs, so finding solutions that meet FedRAMP guidelines is critical.

Section 504:

Section 504 of the Rehabilitation Act of 1973 creates accessibility standards for government agencies and institutions that receive funds from the federal government. This law frequently influences accessibility services in schools, universities and cultural centers like public libraries. However, its standards often overlap with those in the ADA.

Section 508:

Federal government agencies must adhere to Section 508 of the Rehabilitation Act of 1973. That law includes an update regarding online accessibility based on WCAG 2.0 standards. Section 508 creates direct obligations to provide services including captions and audio description. For this reason, government websites face clearer and stricter guidelines than do private businesses.

Speech & Language Processing:

Speech processing refers to the science behind how speech communication works. It focuses on how speech is produced by a speaker and understood by a listener. This process is analyzed and modeled to develop technologies that both produce and understand speech. Language processing works in parallel and explores computational theories of grammar and the meaning conveyed. Language processing encompasses predictions for text, such as automated personal assistants, online searches and more.

SOC 2:

SOC 2 reports provide an assessment of a vendor's security measures. They come into question with regard to speech technologies when businesses use outside vendors and outsource their captioning and transcription efforts where access to customer data is involved. SOC 2 Type I reports list but don't test security, but SOC 2 Type II reports involve a lengthy auditing process and provide a much more comprehensive review of a company's data protection practices. SOC 2 reports let a business know how cautious their vendors are with their information, and how well they are protecting against data breaches.

WCAG:

The Web Content Accessibility Guidelines are considered to be the benchmark for website accessibility. WCAG is part of the web accessibility guidelines published by the Web Accessibility Initiative of the World Wide Web Consortium, the primary international standards organization for the Internet. Many governments and healthcare organizations, among others, have to meet these standards by law. Being WCAG compliant ensures your website and online booking page are accessible to the largest audience possible. In total, there are 61 requirements for WCAG 2.0 and an additional 17 requirements for WCAG 2.1. There are four main categories of accessibility and making video and multimedia accessible with captioning for example is part of these guidelines.

Word Accuracy (WAcc):

WAcc is a metric used to evaluate speech recognition. It considers the number of words recognized correctly from the total number of words spoken. It can be helpful when validating language models. The percent word accuracy is noted as %WAcc = 100 – %WER.

Word error rate (WER):

WER is another metric used to evaluate speech recognition. Accuracy can also be negative. It measures the average number of word errors based on three different error categories: substitution (the correct word is replaced by a different word), insertion (an additional word is guessed that was not stated initially) and deletion (a word was missed). The word error rate is the sum of these errors divided by the number of reference words. The percent word error can be more than 100%.

---

verbit

Even with these definitions on hand, it can be trickier to navigate all of the requirements and choices available to you when considering speech technologies and how to best apply them for greater accessibility efforts. Verbit can serve as an essential partner and guide to you. Our specialized teams and technologies support compliance needs and help to promote more inclusive environments. **Contact us today**.