SPEECH RECOGNITION

Speech recognition, or speech-to-text, is the ability of a machine or program to identify words spoken aloud and convert them into readable text. Rudimentary speech recognition software has a limited vocabulary and may only identify words and phrases when spoken clearly. More sophisticated software can handle natural speech, different accents and various languages.

Speech recognition uses a broad array of research in computer science, linguistics and computer engineering. Many modern devices and text-focused programs have speech recognition functions in them to allow for easier or hands-free use of a device.

Speech recognition and voice recognition are two different technologies and should not be confused:

- **Speech recognition** is used to identify words in spoken language.
- **Voice recognition** is a biometric technology for identifying an individual's voice.

How does speech recognition work?

Speech recognition systems use computer algorithms to process and interpret spoken words and convert them into text. A software program turns the sound a microphone records into written language that computers and humans can understand, following these four steps:

1. analyze the audio;
2. break it into parts;
3. digitize it into a computer-readable format; and
4. use an algorithm to match it to the most suitable text representation.

Speech recognition software must adapt to the highly variable and context-specific nature of human speech. The software algorithms that process and organize audio into text are trained on

different speech patterns, speaking styles, languages, dialects, accents and phrasings. The software also separates spoken audio from background noise that often accompanies the signal.

To meet these requirements, speech recognition systems use two types of models:

- **Acoustic models.** These represent the relationship between linguistic units of speech and audio signals.
- **Language models.** Here, sounds are matched with word sequences to distinguish between words that sound similar.

What applications is speech recognition used for?

Speech recognition systems have quite a few applications. Here is a sampling of them.

Mobile devices. Smartphones use voice commands for call routing, speech-to-text processing, voice dialing and voice search. Users can respond to a text without looking at their devices. On Apple iPhones, speech recognition powers the keyboard and Siri, the virtual assistant. Functionality is available in secondary languages, too. Speech recognition can also be found in word processing applications like Microsoft Word, where users can dictate words to be turned into text.

Education. Speech recognition software is used in language instruction. The software hears the user's speech and offers help with pronunciation.

Customer service. Automated voice assistants listen to customer queries and provides helpful resources.

Healthcare applications. Doctors can use speech recognition software to transcribe notes in real time into healthcare records.

Disability assistance. Speech recognition software can translate spoken words into text using closed captions to enable a person with hearing loss to understand what others are saying.

Speech recognition can also enable those with limited use of their hands to work with computers, using voice commands instead of typing.

Court reporting. Software can be used to transcribe courtroom proceedings, precluding the need for human transcribers.

Emotion recognition. This technology can analyze certain vocal characteristics to determine what emotion the speaker is feeling. Paired with sentiment analysis, this can reveal how someone feels about a product or service.

Hands-free communication. Drivers use voice control for hands-free communication, controlling phones, radios and global positioning systems, for instance.

What are the features of speech recognition systems?

Good speech recognition programs let users customize them to their needs. The features that enable this include:

- **Language weighting.** This feature tells the algorithm to give special attention to certain words, such as those spoken frequently or that are unique to the conversation or subject. For example, the software can be trained to listen for specific product references.
- **Acoustic training.** The software tunes out ambient noise that pollutes spoken audio. Software programs with acoustic training can distinguish speaking style, pace and volume amid the din of many people speaking in an office.
- **Speaker labeling.** This capability enables a program to label individual participants and identify their specific contributions to a conversation.
- **Profanity filtering.** Here, the software filters out undesirable words and language.

What are the different speech recognition algorithms?

The power behind speech recognition features comes from a set of algorithms and technologies. They include the following:

- **Hidden Markov model.** HMMs are used in autonomous systems where a state is partially observable or when all of the information necessary to make a decision is not immediately available to the sensor (in speech recognition's case, a microphone). An example of this is in acoustic modeling, where a program must match linguistic units to audio signals using statistical probability.
- **Natural language processing.** NLP eases and accelerates the speech recognition process.
- **N-grams.** This simple approach to language models creates a probability distribution for a sequence. An example would be an algorithm that looks at the last few words spoken, approximates the history of the sample of speech and uses that to determine the probability of the next word or phrase that will be spoken.
- **Artificial intelligence.** AI and machine learning methods like deep learning and neural networks are common in advanced speech recognition software. These systems use grammar, structure, syntax and composition of audio and voice signals to process speech. Machine learning systems gain knowledge with each use, making them well suited for nuances like accents.

What are the advantages of speech recognition?

There are several advantages to using speech recognition software, including the following:

- **Machine-to-human communication.** The technology enables electronic devices to communicate with humans in natural language or conversational speech.
- **Readily accessible.** This software is frequently installed in computers and mobile devices, making it accessible.
- **Easy to use.** Well-designed software is straightforward to operate and often runs in the background.
- **Continuous, automatic improvement.** Speech recognition systems that incorporate AI become more effective and easier to use over time. As systems complete speech

recognition tasks, they generate more data about human speech and get better at what they do.

What are the disadvantages of speech recognition?

While convenient, speech recognition technology still has a few issues to work through. Limitations include:

- **Inconsistent performance.** The systems may be unable to capture words accurately because of variations in pronunciation, lack of support for some languages and inability to sort through background noise. Ambient noise can be especially challenging. Acoustic training can help filter it out, but these programs aren't perfect. Sometimes it's impossible to isolate the human voice.
- **Speed.** Some speech recognition programs take time to deploy and master. The speech processing may feel relatively slow**.**
- **Source file issues.** Speech recognition success depends on the recording equipment used, not just the software.

Building a Speech Recognizer

Speech Recognition or Automatic Speech Recognition (ASR) is the center of attention for AI projects like robotics. Without ASR, it is not possible to imagine a cognitive robot interacting with a human. However, it is not quite easy to build a speech recognizer.

Difficulties in developing a speech recognition system

Developing a high quality speech recognition system is really a difficult problem. The difficulty of speech recognition technology can be broadly characterized along a number of dimensions as discussed below −

1.Size of the vocabulary − Size of the vocabulary impacts the ease of developing an ASR. Consider the following sizes of vocabulary for a better understanding.

- A small size vocabulary consists of 2-100 words, for example, as in a voice-menu system
- A medium size vocabulary consists of several 100s to 1,000s of words, for example, as in a database-retrieval task

- A large size vocabulary consists of several 10,000s of words, as in a general dictation task.

Note that, the larger the size of vocabulary, the harder it is to perform recognition.

2.Channel characteristics − Channel quality is also an important dimension. For example, human speech contains high bandwidth with full frequency range, while a telephone speech consists of low bandwidth with limited frequency range. Note that it is harder in the latter.

3.Speaking mode − Ease of developing an ASR also depends on the speaking mode, that is whether the speech is in isolated word mode, or connected word mode, or in a continuous speech mode. Note that a continuous speech is harder to recognize.

4.Speaking style − A read speech may be in a formal style, or spontaneous and conversational with casual style. The latter is harder to recognize.

5.Speaker dependency − Speech can be speaker dependent, speaker adaptive, or speaker independent. A speaker independent is the hardest to build.

6.Type of noise − Noise is another factor to consider while developing an ASR. Signal to noise ratio may be in various ranges, depending on the acoustic environment that observes less versus more background noise −

- If the signal to noise ratio is greater than 30dB, it is considered as high range
- If the signal to noise ratio lies between 30dB to 10db, it is considered as medium SNR
- If the signal to noise ratio is lesser than 10dB, it is considered as low range

For example, the type of background noise such as stationary, non-human noise, background speech and crosstalk by other speakers also contributes to the difficulty of the problem.

7.Microphone characteristics − The quality of microphone may be good, average, or below average. Also, the distance between mouth and micro-phone can vary. These factors also should be considered for recognition systems.

Despite these difficulties, researchers worked a lot on various aspects of speech such as understanding the speech signal, the speaker, and identifying the accents.

You will have to follow the steps given below to build a speech recognizer −

Visualizing Audio Signals - Reading from a File and Working on it

This is the first step in building speech recognition system as it gives an understanding of how an audio signal is structured. Some common steps that can be followed to work with audio signals are as follows −

Recording

When you have to read the audio signal from a file, then record it using a microphone, at first.

Sampling

When recording with microphone, the signals are stored in a digitized form. But to work upon it, the machine needs them in the discrete numeric form. Hence, we should perform sampling at a certain frequency and convert the signal into the discrete numerical form. Choosing the high frequency for sampling implies that when humans listen to the signal, they feel it as a continuous audio signal.