Topics: SPEECH RECONGNITION

What is speech recognition?

Speech recognition, also known as automatic speech recognition (ASR), computer speech recognition or speech-to-text, is a capability that enables a program to process human speech into a written format.

While speech recognition is commonly confused with voice recognition, speech recognition focuses on the translation of speech from a verbal format to a text one whereas voice recognition just seeks to identify an individual user’s voice.

While speech technology had a limited vocabulary in the early days, it is utilized in a wide number of industries today, such as automotive, technology, and healthcare. Its adoption has only continued to accelerate in recent years due to advancements in deep learning and big data. Research (link resides outside ibm.com) shows that this market is expected to be worth USD 24.9 billion by 2025.

The best kind of systems also allow organizations to customize and adapt the technology to their specific requirements — everything from language and nuances of speech to brand recognition. For example:

Language weighting: Improve precision by weighting specific words that are spoken frequently (such as product names or industry jargon), beyond terms already in the base vocabulary.

Speaker labeling: Output a transcription that cites or tags each speaker’s contributions to a multi-participant conversation.

Acoustics training: Attend to the acoustical side of the business. Train the system to adapt to an acoustic environment (like the ambient noise in a call center) and speaker styles (like voice pitch, volume and pace).

Speech recognition algorithms

The vagaries of human speech have made development challenging. It’s considered to be one of the most complex areas of computer science – involving linguistics, mathematics and statistics. Speech recognizers are made up of a few components, such as the speech input, feature extraction, feature vectors, a decoder, and a word output. The decoder leverages acoustic models, a pronunciation dictionary, and language models to determine the appropriate output.

Speech recognition technology is evaluated on its accuracy rate, i.e. word error rate (WER), and speed. A number of factors can impact word error rate, such as pronunciation, accent, pitch, volume, and background noise. Reaching human parity – meaning an error rate on par with that of two humans speaking – has long been the goal of speech recognition systems. Research from Lippmann (link resides outside ibm.com) estimates the word error rate to be around 4 percent, but it’s been difficult to replicate the results from this paper.

Various algorithms and computation techniques are used to recognize speech into text and improve the accuracy of transcription.

Natural language processing (NLP): While NLP isn’t necessarily a specific algorithm used in speech recognition, it is the area of artificial intelligence which focuses on the interaction between humans and machines through language through speech and text. Many mobile devices incorporate speech recognition into their systems to conduct voice search—e.g. Siri—or provide more accessibility around texting.

Hidden markov models (HMM): Hidden Markov Models build on the Markov chain model, which stipulates that the probability of a given state hinges on the current state, not its prior states. While a Markov chain model is useful for observable events, such as text inputs, hidden markov models allow us to incorporate hidden events, such as part-of-speech tags, into a probabilistic model. They are utilized as sequence models within speech recognition, assigning labels to each unit—i.e. words, syllables, sentences, etc.—in the sequence. These labels create a mapping with the provided input, allowing it to determine the most appropriate label sequence.

N-grams: This is the simplest type of language model (LM), which assigns probabilities to sentences or phrases. An N-gram is sequence of N-words. For example, “order the pizza” is a trigram or 3-gram and “please order the pizza” is a 4-gram. Grammar and the probability of certain word sequences are used to improve recognition and accuracy.

Neural networks: Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold) and an output. If that output value exceeds a given threshold, it “fires” or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. While neural networks tend to be more accurate and can accept more data, this comes at a performance efficiency cost as they tend to be slower to train compared to traditional language models.

Speaker Diarization (SD): Speaker diarization algorithms identify and segment speech by speaker identity. This helps programs better distinguish individuals in a conversation and is frequently applied at call centers distinguishing customers and sales agents

PROBLEM STATEMENT:

Speech recognition technology is making a huge impact, opening the door to voice-to-text and transcription possibilities. These technologies help deaf and hard of hearing students gain access to higher education which they didn't have previously.

SOLUTION:

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.

DATASETS LINK:

https://paperswithcode.com/datasets?mod=audio