**Existing Voice Recognition Technologies**

Voice recognition will be a key part of the future of communication.

Accurate voice recognition and voice enhancement technology are necessary to provide a reliable voice user interaction experience.

To analyse the sound, it is necessary to frame the sound, i.e. to cut the sound into small sections, each section being called a frame.

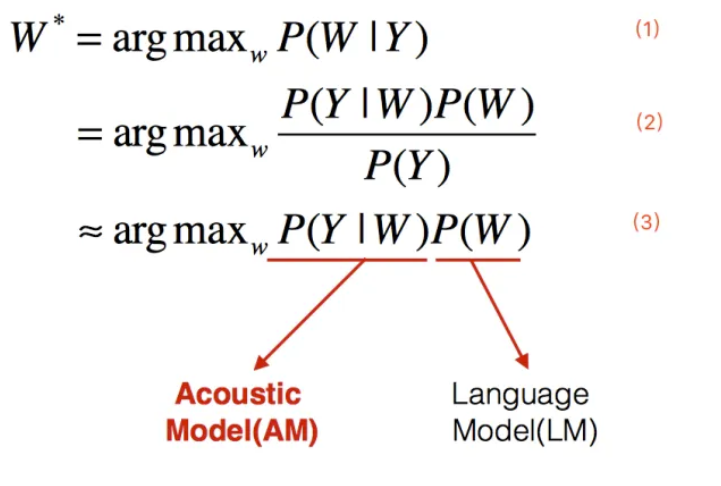
How does speech recognition work? In fact there is no mystery at all, it is nothing more than.

Recognition of frames into states (the hard part).

Combining states into phonemes.

Combining phonemes into words.

Building a state network is a process of unfolding a word-level network into a phoneme network, and then into a state network. The process of speech recognition is in fact a search for the best path in the state network that has the highest probability of corresponding to the speech, which is called "decoding". The algorithm for path search is a dynamic planning pruning algorithm called the Viterbi algorithm, which is used to find the global optimal path.

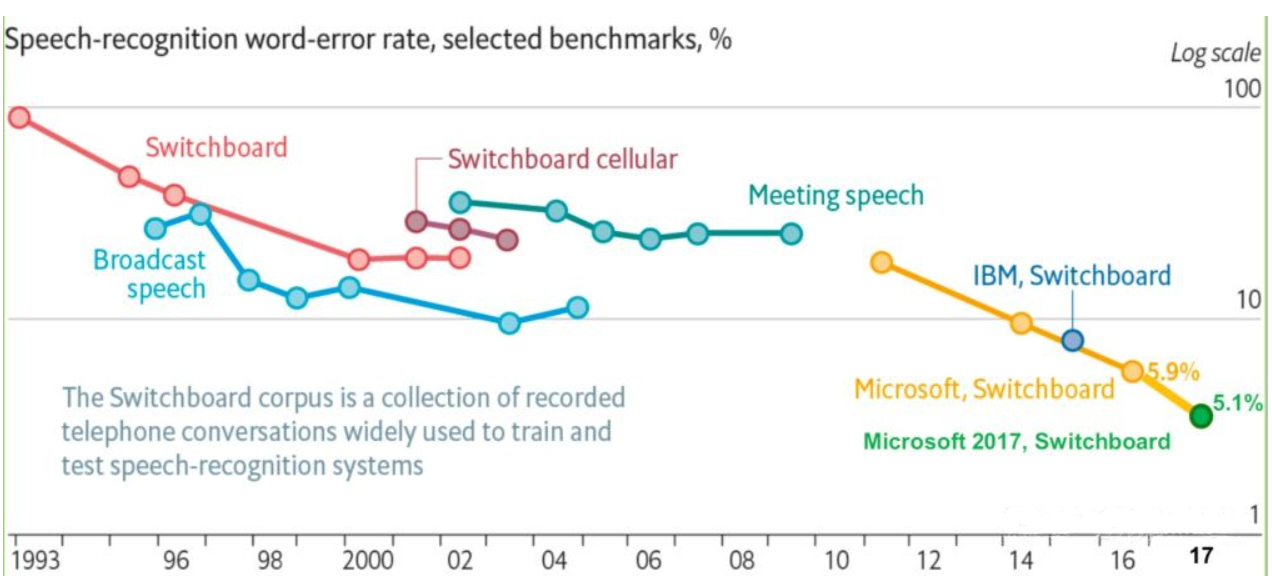


In the above equation W denotes a text sequence and Y denotes a speech input. Equation 1 indicates that the goal of speech recognition is to find the text sequence with the highest probability given the speech input. According to Baye' Rule, Equation 2 can be obtained, where the denominator denotes the probability of occurrence of this audio, which has no parametric relationship compared to the solved text sequence and can be ignored in solving, which in turn leads to Equation 3. The first part of Equation 3 denotes the probability of occurrence of this audio given a text sequence, which is the acoustic model in speech recognition; the second part represents the probability of occurrence of this text sequence, which is the language model in speech recognition.

The accuracy and speed of speech recognition depends on the actual application environment, but the speech recognition rate in quiet environments, standard accents and common vocabulary scenarios has exceeded 95%, implying a speech recognition capability comparable to that of humans, and this is the reason why speech recognition technology is currently developing relatively hot.

With the development of technology, speech recognition in accent, dialect and noise scenarios has now also reached a usable state, especially far-field speech recognition has become one of the most successful technologies used in the global consumer electronics sector with the rise of smart speakers. As voice interaction offers a more natural, convenient and efficient form of communication, voice will certainly become one of the most dominant interfaces for human-computer interaction in the future.

Of course, there are still many shortcomings in the current technology, for example, speech recognition in scenarios such as strong noise, ultra far field, strong interference, multilingualism and large vocabulary still needs to be greatly improved; in addition, multi-person speech recognition and offline speech recognition are also key issues that need to be addressed at present. Although speech recognition cannot be applied to unlimited fields and people, at least we see some hope from the application practice.

History of Voice recognition

**GMM-HMM**

**DNN-HMM**

**END-END**

The mainstream speech recognition framework still consists of 3 components: acoustic model, language model and decoder.

Speech recognition mainly tends to develop in the direction of far-field and fusion, but there are still many difficulties in far-field reliability that have not been broken through, such as multiple rounds of interaction, multi-person noise and other scenarios still to be broken through, as well as the more urgent demand for technologies such as human voice separation. New technologies should solve these problems once and for all, allowing machine hearing to far exceed human perception. This can't just be an algorithmic advance, but requires a common technological upgrade of the entire industry chain, including more advanced sensors and chips with greater arithmetic power.

Machine learning, Artificial Intelligence (AI), and the data that feeds AI are vital factors that will drive improvements in voice recognition.

Machine learning is the lynchpin to voice technology and the ever-growing data that provides AI, making it, and the machines that employ AI, smarter. [AI in voice](https://www.analyticsinsight.net/what-is-the-role-of-voice-automation-ai-and-how-does-it-work/) is built to learn from experiences, identify trends, and provide answers.

<https://www.kardome.com/blog-posts/voice-recognition-technology-challenges-2020-possibilities-future>

[**Artificial intelligence**](https://www.ringcentral.com/gb/en/blog/automated-ai-customer-service-contact-centres/), deep learning, and machine learning are the forces behind speech recognition. Artificial intelligence is used to understand the colloquialisms, abbreviations, and acronyms we use. Machine learning then pieces together the patterns and develops from this data using neural networks.

<https://www.ringcentral.com/gb/en/blog/definitions/voice-recognition/>

* **Automated phone systems**

### ****Google Voice****

### ****Windows Speech Recognition****

