Speech Recognition

Authored by

mohammad looti

June 5, 2026

Speech recognition is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Some speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent".

Speech recognition applications include voice user interfaces such as voice dialing (e.g. Call home"), call routing (e.g. "I would like to make a collect call"), domotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input).

The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process.

From the technology perspective, speech recognition has a long history with several waves of major innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems. These speech industry players include Google, Microsoft, IBM, Baidu, Apple, Amazon, Nuance, SoundHound, IflyTek, CDAC many of which have publicized the core technology in their speech recognition systems as being based on deep learning.

History

Early work

In 1952 three Bell Labs researchers built a system for single-speaker digit recognition. Their system worked by locating the formants in the power spectrum of each utterance. The 1950s era technology was limited to single-speaker systems with vocabularies of around ten words.

Gunnar Fant developed the source-filter model of speech production and published it in 1960,

which proved to be a useful model of speech production.

Unfortunately, funding at Bell Labs dried up for several years when, in 1969, the influential John Pierce wrote an open letter that was critical of speech recognition research. Pierce defunded speech recognition research at Bell Labs where no research on speech recognition was done until Pierce retired and James L. Flanagan took over.

Raj Reddy was the first person to take on continuous speech recognition as a graduate student at Stanford University in the late 1960s. Previous systems required the users to make a pause after each word. Reddy's system was designed to issue spoken commands for the game of chess.

Also around this time Soviet researchers invented the dynamic time warping (DTW) algorithm and used it to create a recognizer capable of operating on a 200-word vocabulary. The DTW algorithm processed the speech signal by dividing it into short frames, e.g. 10ms segments, and processing each frame as a single unit. Although DTW would be superseded by later algorithms, the technique of dividing the signal into frames would carry on. Achieving speaker independence was a major unsolved goal of researchers during this time period.

In 1971, DARPA funded five years of speech recognition research through its Speech Understanding Research program with ambitious end goals including a minimum vocabulary size of 1,000 words. BBN, IBM, Carnegie Mellon and Stanford Research Institute all participated in the program. The government funding revived speech recognition research that had been largely abandoned in the United States after John Pierce's letter.

Despite the fact that CMU's Harpy system met the original goals of the program, many predictions turned out to be nothing more than hype, disappointing DARPA administrators. This disappointment led to DARPA not continuing the funding. Several innovations happened during this time, such as the invention of beam search for use in CMU's Harpy system. The field also benefited from the discovery of several algorithms in other fields such as linear predictive coding and cepstral analysis.

During the late 1960s Leonard Baum developed the mathematics of Markov chains at the Institute for Defense Analysis. At CMU, Raj Reddy's students James Baker and Janet M. Baker began using the Hidden Markov Model (HMM) for speech recognition. James Baker had learned about HMMs from a summer job at the Institute of Defense Analysis during his undergraduate education. The use of HMMs allowed researchers to combine different sources of knowledge, such as acoustics, language, and syntax, in a unified probabilistic model.

Under Fred Jelinek's lead, IBM created a voice activated typewriter called Tangora, which could handle a 20,000 word vocabulary by the mid 1980s. Jelinek's statistical approach put less emphasis on emulating the way the human brain processes and understands speech in favor of

using statistical modeling techniques like HMMs. (Jelinek's group independently discovered the application of HMMs to speech.) This was controversial with linguists since HMMs are too simplistic to account for many common features of human languages. However, the HMM proved to be a highly useful way for modeling speech and replaced dynamic time warping to become the dominant speech recognition algorithm in the 1980s. IBM had a few competitors including Dragon Systems founded by James and Janet M. Baker in 1982. The 1980s also saw the introduction of the n-gram language model. Katz introduced the back-off model in 1987, which allowed language models to use multiple length n-grams. During the same time, also CSELT was using HMM (the diphonies were studied since 1980) to recognize language like Italian. At the same time, CSELT led a series of European projects (Esprit I, II), and summarized the state-of-the-art in a book, later (2013) reprinted.

Much of the progress in the field is owed to the rapidly increasing capabilities of computers. At the end of the DARPA program in 1976, the best computer available to researchers was the PDP-10 with 4 MB ram. Using these computers it could take up to 100 minutes to decode just 30 seconds of speech. A few decades later, researchers had access to tens of thousands of times as much computing power. As the technology advanced and computers got faster, researchers began tackling harder problems such as larger vocabularies, speaker independence, noisy environments and conversational speech. In particular, this shifting to more difficult tasks has characterized DARPA funding of speech recognition since the 1980s. For example, progress was made on speaker independence first by training on a larger variety of speakers and then later by doing explicit speaker adaptation during decoding. Further reductions in word error rate came as researchers shifted acoustic models to be discriminative instead of using maximum likelihood models.

In the mid-Eighties new speech recognition microprocessors were released: for example RIPAC, an independent-speaker recognition (for continuous speech) chip tailored for telephone services, was presented in the Netherlands in 1986. It was designed by CSELT/Elsag and manufactured by SGS.

Practical speech recognition

The 1990s saw the first introduction of commercially successful speech recognition technologies. Two of the earliest products were Dragon Dictate, a consumer product released in 1990 and originally priced at $9,000, and a recognizer from Kurzweil Applied Intelligence released in 1987. AT&T deployed the Voice Recognition Call Processing service in 1992 to route telephone calls without the use of a human operator. The technology was developed by Lawrence Rabiner and others at Bell Labs. By this point, the vocabulary of the typical commercial speech recognition system was larger than the average human vocabulary. Raj Reddy's former student, Xuedong Huang, developed the Sphinx-II system at CMU. The Sphinx-II system was the first to do speaker-

independent, large vocabulary, continuous speech recognition and it had the best performance in DARPA's 1992 evaluation. Handling continuous speech with a large vocabulary was a major milestone in the history of speech recognition. Huang went on to found the speech recognition group at Microsoft in 1993. Raj Reddy's student Kai-Fu Lee joined Apple where, in 1992, he helped develop a speech interface prototype for the Apple computer known as Casper.

Lernout & Hauspie, a Belgium-based speech recognition company, acquired several other companies, including Kurzweil Applied Intelligence in 1997 and Dragon Systems in 2000. The L&H speech technology was used in the Windows XP operating system. L&H was an industry leader until an accounting scandal brought an end to the company in 2001. The speech technology from L&H was bought by ScanSoft which became Nuance in 2005. Apple originally licensed software from Nuance to provide speech recognition capability to its digital assistant Siri.

In the 2000s DARPA sponsored two speech recognition programs: Effective Affordable Reusable Speech-to-Text (EARS) in 2002 and Global Autonomous Language Exploitation (GALE). Four teams participated in the EARS program: IBM, a team led by BBN with LIMSI and Univ. of Pittsburgh, Cambridge University, and a team composed of ISCI, SRI and University of Washington. EARS funded the collection of the Switchboard telephone speech corpus containing 260 hours of recorded conversations from over 500 speakers. The GALE program focused on Arabic and Mandarin broadcast news speech. Google's first effort at speech recognition came in 2007 after hiring some researchers from Nuance. The first product was GOOG-411, a telephone based directory service. The recordings from GOOG-411 produced valuable data that helped Google improve their recognition systems. Google voice search is now supported in over 30 languages.

In the United States, the National Security Agency has made use of a type of speech recognition for keyword spotting since at least 2006. This technology allows analysts to search through large volumes of recorded conversations and isolate mentions of keywords. Recordings can be indexed and analysts can run queries over the database to find conversations of interest. Some government research programs focused on intelligence applications of speech recognition, e.g. DARPA's EARS's program and IARPA's Babel program.

Modern systems

In the early 2000s, speech recognition was still dominated by traditional approaches such as Hidden Markov Models combined with feedforward artificial neural networks. Today, however, many aspects of speech recognition have been taken over by a deep learning method called Long short-term memory (LSTM), a recurrent neural network published by Sepp Hochreiter & Jürgen Schmidhuber in 1997. LSTM RNNs avoid the vanishing gradient problem and can learn "Very Deep Learning" tasks that require memories of events that happened thousands of discrete time

steps ago, which is important for speech. Around 2007, LSTM trained by Connectionist Temporal Classification (CTC) started to outperform traditional speech recognition in certain applications. In 2015, Google's speech recognition reportedly experienced a dramatic performance jump of 49% through CTC-trained LSTM, which is now available through Google Voice to all smartphone users.

The use of deep feedforward (non-recurrent) networks for acoustic modeling was introduced during later part of 2009 by Geoffrey Hinton and his students at University of Toronto and by Li Deng and colleagues at Microsoft Research, initially in the collaborative work between Microsoft and University of Toronto which was subsequently expanded to include IBM and Google (hence "The shared views of four research groups" subtitle in their 2012 review paper). A Microsoft research executive called this innovation "the most dramatic change in accuracy since 1979." In contrast to the steady incremental improvements of the past few decades, the application of deep learning decreased word error rate by 30%. This innovation was quickly adopted across the field. Researchers have begun to use deep learning techniques for language modeling as well.

In the long history of speech recognition, both shallow form and deep form (e.g. recurrent nets) of artificial neural networks had been explored for many years during 1980s, 1990s and a few years into the 2000s. But these methods never won over the non-uniform internal-handcrafting Gaussian mixture model/Hidden Markov model (GMM-HMM) technology based on generative models of speech trained discriminatively. A number of key difficulties had been methodologically analyzed in the 1990s, including gradient diminishing and weak temporal correlation structure in the neural predictive models. All these difficulties were in addition to the lack of big training data and big computing power in these early days. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning starting around 2009-2010 that had overcome all these difficulties. Hinton et al. and Deng et al. reviewed part of this recent history about how their collaboration with each other and then with colleagues across four groups (University of Toronto, Microsoft, Google, and IBM) ignited a renaissance of applications of deep feedforward neural networks to speech recognition.

Models, methods, and algorithms

Both acoustic modeling and language modeling are important parts of modern statistically-based speech recognition algorithms. Hidden Markov models (HMMs) are widely used in many systems. Language modeling is also used in many other natural language processing applications such as document classification or statistical machine translation.

Hidden Markov models

Modern general-purpose speech recognition systems are based on Hidden Markov Models. These

are statistical models that output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scale (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes.

Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of n-dimensional real-valued vectors (with n being a small integer, such as 10), outputting one of these every 10 milliseconds. The vectors would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians, which will give a likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phonemes (so phonemes with different left and right context have different realizations as HMM states); it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization (VTLN) for male-female normalization and maximum likelihood linear regression (MLLR) for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis (HLDA); or might skip the delta and delta-delta coefficients and use splicing and an LDA-based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semi-tied co variance transform (also known as maximum likelihood linear transform, or MLLT). Many systems use so-called discriminative training techniques that dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Examples are maximum mutual information (MMI), minimum classification error (MCE) and minimum phone error (MPE).

Decoding of the speech (the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence) would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model, which includes both the acoustic and language model information, and combining it statically beforehand (the finite state transducer, or FST, approach).

A possible improvement to decoding is to keep a set of good candidates instead of just keeping the best candidate, and to use a better scoring function (re scoring) to rate these good candidates so that we may pick the best one according to this refined score. The set of candidates can be kept either as a list (the N-best list approach) or as a subset of the models (a lattice). Re scoring is usually done by trying to minimize the Bayes risk (or an approximation thereof): Instead of taking the source sentence with maximal probability, we try to take the sentence that minimizes the expectancy of a given loss function with regards to all possible transcriptions (i.e., we take the sentence that minimizes the average distance to other possible sentences weighted by their estimated probability). The loss function is usually the Levenshtein distance, though it can be different distances for specific tasks; the set of possible transcriptions is, of course, pruned to maintain tractability. Efficient algorithms have been devised to re score lattices represented as weighted finite state transducers with edit distances represented themselves as a finite state transducer verifying certain assumptions.

Dynamic time warping (DTW)-based speech recognition

Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach.

Dynamic time warping is an algorithm for measuring similarity between two sequences that may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another he or she were walking more quickly, or even if there were accelerations and deceleration during the course of one observation. DTW has been applied to video, audio, and graphics - indeed, any data that can be turned into a linear representation can be analyzed with DTW.

A well-known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g., time series) with certain restrictions. That is, the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models.

Neural networks

Neural networks emerged as an attractive acoustic modeling approach in ASR in the late 1980s. Since then, neural networks have been used in many aspects of speech recognition such as phoneme classification, isolated word recognition, audiovisual speech recognition, audiovisual speaker recognition and speaker adaptation.

In contrast to HMMs, neural networks make no assumptions about feature statistical properties and

have several qualities making them attractive recognition models for speech recognition. When used to estimate the probabilities of a speech feature segment, neural networks allow discriminative training in a natural and efficient manner. Few assumptions on the statistics of input features are made with neural networks. However, in spite of their effectiveness in classifying short-time units such as individual phones and isolated words, neural networks are rarely successful for continuous recognition tasks, largely because of their lack of ability to model temporal dependencies.

However, recently LSTM Recurrent Neural Networks (RNNs) and Time Delay Neural Networks(TDNN's) have been used which have been shown to be able to identify latent temporal dependencies and use this information to perform the task of speech recognition.

Deep Neural Networks and Denoising Autoencoders were also being experimented with to tackle this problem in an effective manner.

Due to the inability of feedforward Neural Networks to model temporal dependencies, an alternative approach is to use neural networks as a pre-processing e.g. feature transformation, dimensionality reduction, for the HMM based recognition.

Deep feedforward and recurrent neural networks

A deep feedforward neural network (DNN) is an artificial neural network with multiple hidden layers of units between the input and output layers. Similar to shallow neural networks, DNNs can model complex non-linear relationships. DNN architectures generate compositional models, where extra layers enable composition of features from lower layers, giving a huge learning capacity and thus the potential of modeling complex patterns of speech data.

A success of DNNs in large vocabulary speech recognition occurred in 2010 by industrial researchers, in collaboration with academic researchers, where large output layers of the DNN based on context dependent HMM states constructed by decision trees were adopted. See comprehensive reviews of this development and of the state of the art as of October 2014 in the recent Springer book from Microsoft Research. See also the related background of automatic speech recognition and the impact of various machine learning paradigms including notably deep learning in recent overview articles.

One fundamental principle of deep learning is to do away with hand-crafted feature engineering and to use raw features. This principle was first explored successfully in the architecture of deep autoencoder on the "raw" spectrogram or linear filter-bank features, showing its superiority over the Mel-Cepstral features which contain a few stages of fixed transformation from spectrograms. The true "raw" features of speech, waveforms, have more recently been shown to produce excellent larger-scale speech recognition results.

End-to-end automatic speech recognition

Since 2014, there has been much research interest in "end-to-end" ASR. Traditional phonetic-based (i.e., all HMM-based model) approaches required separate components and training for the pronunciation, acoustic and language model. End-to-end models jointly learn all the components of the speech recognizer. This is valuable since it simplifies the training process and deployment process. For example, a n-gram language model is required for all HMM-based systems, and a typical n-gram language model often takes several gigabytes in memory making them impractical to deploy on mobile devices. Consequently, modern commercial ASR systems from Google and Apple (as of 2017) are deployed on the cloud and require a network connection as opposed to the device locally.

The first attempt of end-to-end ASR was with Connectionist Temporal Classification (CTC) based systems introduced by Alex Graves of Google DeepMind and Navdeep Jaitly of the University of Toronto in 2014. The model consisted of recurrent neural networks and a CTC layer. Jointly, the RNN-CTC model learns the pronunciation and acoustic model together, however it is incapable of learning the language due to conditional independence assumptions similar to a HMM. Consequently, CTC models can directly learn to map speech acoustics to English characters, but the models make many common spelling mistakes and must rely on a separate language model to clean up the transcripts. Later, Baidu expanded on the work with extremely large datasets and demonstrated some commercial success in Chinese Mandarin and English. In 2016, University of Oxford presented LipNet, the first end-to-end sentence-level lip reading model, using spatiotemporal convolutions coupled with an RNN-CTC architecture, surpassing human-level performance in a restricted grammar dataset.

An alternative approach to CTC-based models are attention-based models. Attention-based ASR models were introduced simultaneously by Chan et al. of Carnegie Mellon University and Google Brain and Bahdanaua et al. of the University of Montreal in 2016. The model named "Listen, Attend and Spell" (LAS), literally "listens" to the acoustic signal, pays "attention" to different parts of the signal and "spells" out the transcript one character at a time. Unlike CTC-based models, attention-based models do not have conditional-independence assumptions and can learn all the components of a speech recognizer including the pronunciation, acoustic and language model directly. This means, during deployment, there is no need to carry around a language model making it very practical for deployment onto applications with limited memory. By the end of 2016, the attention-based models have seen considerable success including outperforming the CTC models (with or without an external language model). Various extensions have been proposed since the original LAS model. Latent Sequence Decompositions (LSD) was proposed by Carnegie Mellon University, MIT and Google Brain to directly emit sub-word units which are more natural than English characters; University of Oxford and Google DeepMind extended LAS to Watch, Listen, Attend and Spell" (WLAS) to handle lip reading surpassing human-level performance.

Applications

In-car systems

Typically a manual control input, for example by means of a finger control on the steering-wheel, enables the speech recognition system and this is signalled to the driver by an audio prompt. Following the audio prompt, the system has a "listening window" during which it may accept a speech input for recognition.

Simple voice commands may be used to initiate phone calls, select radio stations or play music from a compatible smartphone, MP3 player or music-loaded flash drive. Voice recognition capabilities vary between car make and model. Some of the most recent car models offer natural-language speech recognition in place of a fixed set of commands, allowing the driver to use full sentences and common phrases. With such systems there is, therefore, no need for the user to memorize a set of fixed command words.

Health care

Medical documentation

In the health care sector, speech recognition can be implemented in front-end or back-end of the medical documentation process. Front-end speech recognition is where the provider dictates into a speech-recognition engine, the recognized words are displayed as they are spoken, and the dictator is responsible for editing and signing off on the document. Back-end or deferred speech recognition is where the provider dictates into a digital dictation system, the voice is routed through a speech-recognition machine and the recognized draft document is routed along with the original voice file to the editor, where the draft is edited and report finalized. Deferred speech recognition is widely used in the industry currently.

One of the major issues relating to the use of speech recognition in healthcare is that the American Recovery and Reinvestment Act of 2009 (ARRA) provides for substantial financial benefits to physicians who utilize an EMR according to "Meaningful Use" standards. These standards require that a substantial amount of data be maintained by the EMR (now more commonly referred to as an Electronic Health Record or EHR). The use of speech recognition is more naturally suited to the generation of narrative text, as part of a radiology/pathology interpretation, progress note or discharge summary: the ergonomic gains of using speech recognition to enter structured discrete data (e.g., numeric values or codes from a list or a controlled vocabulary) are relatively minimal for people who are sighted and who can operate a keyboard and mouse.

A more significant issue is that most EHRs have not been expressly tailored to take advantage of voice-recognition capabilities. A large part of the clinician's interaction with the EHR involves

navigation through the user interface using menus, and tab/button clicks, and is heavily dependent on keyboard and mouse: voice-based navigation provides only modest ergonomic benefits. By contrast, many highly customized systems for radiology or pathology dictation implement voice "macros", where the use of certain phrases - e.g., "normal report", will automatically fill in a large number of default values and/or generate boilerplate, which will vary with the type of the exam - e.g., a chest X-ray vs. a gastrointestinal contrast series for a radiology system.

As an alternative to this navigation by hand, cascaded use of speech recognition and information extraction has been studied as a way to fill out a handover form for clinical proofing and sign-off. The results are encouraging, and the paper also opens data, together with the related performance benchmarks and some processing software, to the research and development community for studying clinical documentation and language-processing.

Therapeutic use

Prolonged use of speech recognition software in conjunction with word processors has shown benefits to short-term-memory restrengthening in brain AVM patients who have been treated with resection. Further research needs to be conducted to determine cognitive benefits for individuals whose AVMs have been treated using radiologic techniques.

Military

High-performance fighter aircraft

Substantial efforts have been devoted in the last decade to the test and evaluation of speech recognition in fighter aircraft. Of particular note have been the US program in speech recognition for the Advanced Fighter Technology Integration (AFTI)/F-16 aircraft (F-16 VISTA), the program in France for Mirage aircraft, and other programs in the UK dealing with a variety of aircraft platforms. In these programs, speech recognizers have been operated successfully in fighter aircraft, with applications including: setting radio frequencies, commanding an autopilot system, setting steer-point coordinates and weapons release parameters, and controlling flight display.

Working with Swedish pilots flying in the JAS-39 Gripen cockpit, Englund (2004) found recognition deteriorated with increasing g-loads. The report also concluded that adaptation greatly improved the results in all cases and that the introduction of models for breathing was shown to improve recognition scores significantly. Contrary to what might have been expected, no effects of the broken English of the speakers were found. It was evident that spontaneous speech caused problems for the recognizer, as might have been expected. A restricted vocabulary, and above all, a proper syntax, could thus be expected to improve recognition accuracy substantially.

The Eurofighter Typhoon, currently in service with the UK RAF, employs a speaker-dependent system, requiring each pilot to create a template. The system is not used for any safety-critical or weapon-critical tasks, such as weapon release or lowering of the undercarriage, but is used for a wide range of other cockpit functions. Voice commands are confirmed by visual and/or aural feedback. The system is seen as a major design feature in the reduction of pilot workload, and even allows the pilot to assign targets to his aircraft with two simple voice commands or to any of his wingmen with only five commands.

Speaker-independent systems are also being developed and are under test for the F35 Lightning II (JSF) and the Alenia Aermacchi M-346 Master lead-in fighter trainer. These systems have produced word accuracy scores in excess of 98%.

Helicopters

The problems of achieving high recognition accuracy under stress and noise pertain strongly to the helicopter environment as well as to the jet fighter environment. The acoustic noise problem is actually more severe in the helicopter environment, not only because of the high noise levels but also because the helicopter pilot, in general, does not wear a facemask, which would reduce acoustic noise in the microphone. Substantial test and evaluation programs have been carried out in the past decade in speech recognition systems applications in helicopters, notably by the U.S. Army Avionics Research and Development Activity (AVRADA) and by the Royal Aerospace Establishment (RAE) in the UK. Work in France has included speech recognition in the Puma helicopter. There has also been much useful work in Canada. Results have been encouraging, and voice applications have included: control of communication radios, setting of navigation systems, and control of an automated target handover system.

As in fighter applications, the overriding issue for voice in helicopters is the impact on pilot effectiveness. Encouraging results are reported for the AVRADA tests, although these represent only a feasibility demonstration in a test environment. Much remains to be done both in speech recognition and in overall speech technology in order to consistently achieve performance improvements in operational settings.

Training air traffic controllers

Training for air traffic controllers (ATC) represents an excellent application for speech recognition systems. Many ATC training systems currently require a person to act as a "pseudo-pilot", engaging in a voice dialog with the trainee controller, which simulates the dialog that the controller would have to conduct with pilots in a real ATC situation. Speech recognition and synthesis techniques offer the potential to eliminate the need for a person to act as pseudo-pilot, thus reducing training and support personnel. In theory, Air controller tasks are also characterized by

highly structured speech as the primary output of the controller, hence reducing the difficulty of the speech recognition task should be possible. In practice, this is rarely the case. The FAA document 7110.65 details the phrases that should be used by air traffic controllers. While this document gives less than 150 examples of such phrases, the number of phrases supported by one of the simulation vendors speech recognition systems is in excess of 500,000.

The USAF, USMC, US Army, US Navy, and FAA as well as a number of international ATC training organizations such as the Royal Australian Air Force and Civil Aviation Authorities in Italy, Brazil, and Canada are currently using ATC simulators with speech recognition from a number of different vendors.

Telephony and other domains

ASR is now commonplace In the field of telephony, and is becoming more widespread in the field of computer gaming and simulation. Despite the high level of integration with word processing in general personal computing. However, ASR in the field of document production has not seen the expected increases in use.

The improvement of mobile processor speeds has made speech recognition practical in smartphones. Speech is used mostly as a part of a user interface, for creating predefined or custom speech commands. Leading software vendors in this field are: Google, Microsoft Corporation (Microsoft Voice Command), Digital Syphon (Sonic Extractor), LumenVox, Nuance Communications (Nuance Voice Control), Voci Technologies, VoiceBox Technology, Speech Technology Center, Vito Technologies (VITO Voice2Go), Speereo Software (Speereo Voice Translator), Verbyx VRX and SVOX.

Usage in education and daily life

For language learning, speech recognition can be useful for learning a second language. It can teach proper pronunciation, in addition to helping a person develop fluency with their speaking skills.

Students who are blind (see Blindness and education) or have very low vision can benefit from using the technology to convey words and then hear the computer recite them, as well as use a computer by commanding with their voice, instead of having to look at the screen and keyboard.

Students who are physically disabled or suffer from Repetitive strain injury/other injuries to the upper extremities can be relieved from having to worry about handwriting, typing, or working with scribe on school assignments by using speech-to-text programs. They can also utilize speech recognition technology to freely enjoy searching the Internet or using a computer at home without

having to physically operate a mouse and keyboard.

Speech recognition can allow students with learning disabilities to become better writers. By saying the words aloud, they can increase the fluidity of their writing, and be alleviated of concerns regarding spelling, punctuation, and other mechanics of writing. Also, see Learning disability.

Use of voice recognition software, in conjunction with a digital audio recorder and a personal computer running word-processing software has proven to be positive for restoring damaged short-term-memory capacity, in stroke and craniotomy individuals.

People with disabilities

People with disabilities can benefit from speech recognition programs. For individuals that are Deaf or Hard of Hearing, speech recognition software is used to automatically generate a closed-captioning of conversations such as discussions in conference rooms, classroom lectures, and/or religious services.

Speech recognition is also very useful for people who have difficulty using their hands, ranging from mild repetitive stress injuries to involve disabilities that preclude using conventional computer input devices. In fact, people who used the keyboard a lot and developed RSI became an urgent early market for speech recognition. Speech recognition is used in deaf telephony, such as voicemail to text, relay services, and captioned telephone. Individuals with learning disabilities who have problems with thought-to-paper communication (essentially they think of an idea but it is processed incorrectly causing it to end up differently on paper) can possibly benefit from the software but the technology is not bug proof. Also the whole idea of speak to text can be hard for intellectually disabled person's due to the fact that it is rare that anyone tries to learn the technology to teach the person with the disability.

This type of technology can help those with dyslexia but other disabilities are still in question. The effectiveness of the product is the problem that is hindering it being effective. Although a kid may be able to say a word depending on how clear they say it the technology may think they are saying another word and input the wrong one. Giving them more work to fix, causing them to have to take more time with fixing the wrong word.

Further applications

Aerospace (e.g. space exploration, spacecraft, etc.) NASA's Mars Polar Lander used speech recognition technology from Sensory, Inc. in the Mars Microphone on the Lander
Automatic subtitling with speech recognition
Automatic translation

Court reporting (Realtime Speech Writing)
eDiscovery (Legal discovery)
Hands-free computing: Speech recognition computer user interface
Home automation
Interactive voice response
Mobile telephony, including mobile email
Multimodal interaction
Pronunciation evaluation in computer-aided language learning applications
Robotics
Speech-to-text reporter (transcription of speech into text, video captioning, Court reporting)
Telematics (e.g. vehicle Navigation Systems)
Transcription (digital speech-to-text)
Video games, with Tom Clancy's EndWar and Lifeline as working examples
Virtual assistant (e.g. Apple's Siri)

Performance

The performance of speech recognition systems is usually evaluated in terms of accuracy and speed. Accuracy is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

Speech recognition by machine is a very complex problem, however. Vocalizations vary in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed. Speech is distorted by a background noise and echoes, electrical characteristics. Accuracy of speech recognition may vary with the following:

Vocabulary size and confusability
Speaker dependence versus independence
Isolated, discontinuous or continuous speech
Task and language constraints
Read versus spontaneous speech
Adverse conditions

Accuracy

As mentioned earlier in this article, accuracy of speech recognition may vary depending on the following factors:

Error rates increase as the vocabulary size grows:

e.g. the 10 digits "zero" to "nine" can be recognized essentially perfectly, but vocabulary sizes of

200, 5000 or 100000 may have error rates of 3%, 7% or 45% respectively.

Vocabulary is hard to recognize if it contains confusable words:

e.g. the 26 letters of the English alphabet are difficult to discriminate because they are confusable words (most notoriously, the E-set: "B, C, D, E, G, P, T, V, Z"); an 8% error rate is considered good for this vocabulary.

Speaker dependence vs. independence:

A speaker-dependent system is intended for use by a single speaker.

A speaker-independent system is intended for use by any speaker (more difficult).

Isolated, Discontinuous or continuous speech

With isolated speech, single words are used, therefore it becomes easier to recognize the speech.

With discontinuous speech full sentences separated by silence are used, therefore it becomes easier to recognize the speech as well as with isolated speech.

With continuous speech naturally spoken sentences are used, therefore it becomes harder to recognize the speech, different from both isolated and discontinuous speech.

Task and language constraints

e.g. Querying application may dismiss the hypothesis "The apple is red."

e.g. Constraints may be semantic; rejecting "The apple is angry."

e.g. Syntactic; rejecting "Red is apple the."

Constraints are often represented by a grammar.

Read vs. Spontaneous Speech

When a person reads it's usually in a context that has been previously prepared, but when a person uses spontaneous speech, it is difficult to recognize the speech because of the disfluencies (like "uh" and "um", false starts, incomplete sentences, stuttering, coughing, and laughter) and limited vocabulary.

Environmental noise (e.g. Noise in a car or a factory)

Acoustical distortions (e.g. echoes, room acoustics)

Speech recognition is a multi-levelled pattern recognition task.

Acoustical signals are structured into a hierarchy of units;

e.g. Phonemes, Words, Phrases, and Sentences;

Each level provides additional constraints;

e.g. Known word pronunciations or legal word sequences, which can compensate for errors or uncertainties at lower level;

This hierarchy of constraints are exploited;

By combining decisions probabilistically at all lower levels, and making more deterministic decisions only at the highest level, speech recognition by a machine is a process broken into several phases. Computationally, it is a problem in which a sound pattern has to be recognized or classified into a category that represents a meaning to a human. Every acoustic signal can be broken in smaller more basic sub-signals. As the more complex sound signal is broken into the smaller sub-sounds, different levels are created, where at the top level we have complex sounds, which are made of simpler sounds on lower level, and going to lower levels even more, we create more basic and shorter and simpler sounds. The lowest level, where the sounds are the most fundamental, a machine would check for simple and more probabilistic rules of what sound should represent. Once these sounds are put together into more complex sound on upper level, a new set of more deterministic rules should predict what new complex sound should represent. The most upper level of a deterministic rule should figure out the meaning of complex expressions. In order to expand our knowledge about speech recognition we need to take into a consideration neural networks. There are four steps of neural network approaches:

Digitize the speech that we want to recognize

For telephone speech the sampling rate is 8000 samples per second;

Compute features of spectral-domain of the speech (with Fourier transform);

computed every 10 ms, with one 10 ms section called a frame;

Analysis of four-step neural network approaches can be explained by further information. Sound is produced by air (or some other medium) vibration, which we register by ears, but machines by receivers. Basic sound creates a wave which has 2 descriptions; Amplitude (how strong is it), and frequency (how often it vibrates per second).

The sound waves can be digitized: Sample a strength at short intervals like in picture above to get bunch of numbers that approximate at each time step the strength of a wave. Collection of these

numbers represent analog wave. This new wave is digital. Sound waves are complicated because they superimpose one on top of each other. Like the waves would. This way they create odd-looking waves. For example, if there are two waves that interact with each other we can add them which creates new odd-looking wave.

Neural network classifies features into phonetic-based categories;

Given basic sound blocks that a machine digitized, one has a bunch of numbers which describe a wave and waves describe words. Each frame has a unit block of sound, which are broken into basic sound waves and represented by numbers which, after Fourier Transform, can be statistically evaluated to set to which class of sounds it belongs. The nodes in the figure on a slide represent a feature of a sound in which a feature of a wave from the first layer of nodes to the second layer of nodes based on statistical analysis. This analysis depends on programmer's instructions. At this point, a second layer of nodes represents higher level features of a sound input which is again statistically evaluated to see what class they belong to. Last level of nodes should be output nodes that tell us with high probability what original sound really was.

Search to match the neural-network output scores for the best word, to determine the word that was most likely uttered.

Security concerns

Speech recognition can become a means of attack, theft, or accidental operation. For example, activation words like "Alexa" spoken in an audio or video broadcast or by non-owners in the same room can cause devices in audience homes and offices to start listening for input inappropriately, or possibly take an unwanted action. Another demonstrated approach is to transmit ultrasound and attempt to send commands without nearby people noticing.

Further information

Conferences and journals

Popular speech recognition conferences held each year or two include SpeechTEK and SpeechTEK Europe, ICASSP, Interspeech/Eurospeech, and the IEEE ASRU. Conferences in the field of natural language processing, such as ACL, NAACL, EMNLP, and HLT, are beginning to include papers on speech processing. Important journals include the IEEE Transactions on Speech and Audio Processing (later renamed IEEE Transactions on Audio, Speech and Language Processing and since Sept 2014 renamed IEEE/ACM Transactions on Audio, Speech and Language Processing--after merging with an ACM publication), Computer Speech and Language, and Speech Communication.

Books

Books like "Fundamentals of Speech Recognition" by Lawrence Rabiner can be useful to acquire basic knowledge but may not be fully up to date (1993). Another good source can be "Statistical Methods for Speech Recognition" by Frederick Jelinek and "Spoken Language Processing (2001)" by Xuedong Huang etc. More up to date are Computer Speech", by Manfred R. Schroeder, second edition published in 2004, and "Speech Processing: A Dynamic and Optimization-Oriented Approach" published in 2003 by Li Deng and Doug O'Shaughnessey. The recently updated textbook of "Speech and Language Processing (2008)" by Jurafsky and Martin presents the basics and the state of the art for ASR. Speaker recognition also uses the same features, most of the same front-end processing, and classification techniques as is done in speech recognition. A most recent comprehensive textbook, "Fundamentals of Speaker Recognition" is an in depth source for up to date details on the theory and practice. A good insight into the techniques used in the best modern systems can be gained by paying attention to government sponsored evaluations such as those organised by DARPA (the largest speech recognition-related project ongoing as of 2007 is the GALE project, which involves both speech recognition and translation components).

A good and accessible introduction to speech recognition technology and its history is provided by the general audience book "The Voice in the Machine. Building Computers That Understand Speech" by Roberto Pieraccini (2012).

The most recent book on speech recognition is "Automatic Speech Recognition: A Deep Learning Approach" (Publisher: Springer) written by D. Yu and L. Deng published near the end of 2014, with highly mathematically-oriented technical detail on how deep learning methods are derived and implemented in modern speech recognition systems based on DNNs and related deep learning methods. A related book, published earlier in 2014, "Deep Learning: Methods and Applications" by L. Deng and D. Yu provides a less technical but more methodology-focused overview of DNN-based speech recognition during 2009-2014, placed within the more general context of deep learning applications including not only speech recognition but also image recognition, natural language processing, information retrieval, multimodal processing, and multitask learning.

Software

In terms of freely available resources, Carnegie Mellon University's Sphinx toolkit is one place to start to both learn about speech recognition and to start experimenting. Another resource (free but copyrighted) is the HTK book (and the accompanying HTK toolkit). For more recent and state-of-the-art techniques, Kaldi toolkit can be used.

A Demo of an on-line speech recognizer is available on Cobalt's webpage.