WIKIPEDIA
The Free Encyclopedia

WIKIPEDIA

Speech recognition

Speech recognition is an interdisciplinary subfield of computer science and computational linguistics that develops methodologies and technologies that enable the recognition and translation of spoken language into text by computers. It is also known as **automatic speech recognition** (**ASR**), **computer speech recognition** or **speech-to-text** (**STT**). It incorporates knowledge and research in the computer science, linguistics and computer engineering fields. The reverse process is speech synthesis.

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"[1] 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 key words (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), determining speaker characteristics,[2] speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed direct voice input). Automatic pronunciation assessment is used in education such as for spoken language learning.

The term *voice recognition*[3][4][5] or *speaker identification*[6][7][8] 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.

History

The key areas of growth were: vocabulary size, speaker independence, and processing speed.

Pre-1970

- **1952** – Three Bell Labs researchers, Stephen Balashek,[9] R. Biddulph, and K. H. Davis built a system called "Audrey"[10] for single-speaker digit recognition. Their system located the formants in the power spectrum of each utterance.[11]
- **1960** – Gunnar Fant developed and published the source-filter model of speech production.
- **1962** – IBM demonstrated its 16-word "Shoebox" machine's speech recognition capability at the 1962 World's Fair.[12]
- **1966** – Linear predictive coding (LPC), a speech coding method, was first proposed by Fumitada Itakura of Nagoya University and Shuzo Saito of Nippon Telegraph and Telephone (NTT), while working on speech recognition.[13]
- **1969** – Funding at Bell Labs dried up for several years when, in 1969, the influential John Pierce wrote an open letter that was critical of and defunded speech recognition research.[14] This defunding lasted 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 users to pause after each word. Reddy's system issued spoken commands for playing chess.

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.[15] DTW processed speech 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 carried on. Achieving speaker independence remained unsolved at this time period.

1970–1990

- **1971** – DARPA funded five years for *Speech Understanding Research*, speech recognition research seeking a minimum vocabulary size of 1,000 words. They thought speech *understanding* would be key to making progress in speech *recognition*, but this later proved untrue.[16] BBN, IBM, Carnegie Mellon and Stanford Research Institute all participated in the program.[17][18] This revived speech recognition research post John Pierce's letter.

- **1972** – The IEEE Acoustics, Speech, and Signal Processing group held a conference in Newton, Massachusetts.
- **1976** – The first ICASSP was held in Philadelphia, which since then has been a major venue for the publication of research on speech recognition.[19]

During the late 1960s Leonard Baum developed the mathematics of Markov chains at the Institute for Defense Analysis. A decade later, at CMU, Raj Reddy's students James Baker and Janet M. Baker began using the hidden Markov model (HMM) for speech recognition.[20] James Baker had learned about HMMs from a summer job at the Institute of Defense Analysis during his undergraduate education.[21] The use of HMMs allowed researchers to combine different sources of knowledge, such as acoustics, language, and syntax, in a unified probabilistic model.

- By the **mid-1980s** IBM's Fred Jelinek's team created a voice activated typewriter called Tangora, which could handle a 20,000-word vocabulary[22] 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.[21]) This was controversial with linguists since HMMs are too simplistic to account for many common features of human languages.[23] 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.[24][25]

- **1982** – Dragon Systems, founded by James and Janet M. Baker,[26] was one of IBM's few competitors.

Practical speech recognition

The 1980s also saw the introduction of the n-gram language model.

- **1987** – The back-off model allowed language models to use multiple length n-grams, and CSELT[27] used HMM to recognize languages (both in software and in hardware specialized processors, e.g. RIPAC).

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.[28] It could take up to 100 minutes to decode just 30 seconds of speech.[29]

Two practical products were:

- **1984** – was released the Apricot Portable with up to 4096 words support, of which only 64 could be held in RAM at a time.[30]
- **1987** – a recognizer from Kurzweil Applied Intelligence
- **1990** – Dragon Dictate, a consumer product released in 1990[31][32] AT&T deployed the Voice Recognition Call Processing service in 1992 to route telephone calls without the use of a human operator.[33] 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.[28] 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.[34]

2000s

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 ICSI, 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.[35] 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.[36] 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.[37] 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.

In the early 2000s, speech recognition was still dominated by traditional approaches such as hidden Markov models combined with feedforward artificial neural networks.[38] 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.[39] LSTM RNNs avoid the vanishing gradient problem and can learn "Very Deep Learning" tasks[40] 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)[41] started to outperform traditional speech recognition in certain applications.[42] 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.[43] Transformers, a type of neural network based solely on "attention", have been widely adopted in computer vision[44][45] and language modeling,[46][47] sparking the interest of adapting such models to new domains, including speech recognition.[48][49][50] Some recent papers reported superior performance levels using transformer models for speech recognition, but these models usually require large scale training datasets to reach high performance levels.

The use of deep feedforward (non-recurrent) networks for acoustic modeling was introduced during the later part of 2009 by Geoffrey Hinton and his students at the University of Toronto and by Li Deng[51] and colleagues at Microsoft Research, initially in the collaborative work between Microsoft and the 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).[52][53][54] A Microsoft research executive called this innovation "the most dramatic change in accuracy since 1979".[55] In contrast to the steady incremental improvements of the past few decades, the application of deep learning decreased word error rate by 30%.[55] 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.[56][57][58] 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.[59] A number of key difficulties had been methodologically analyzed in the 1990s, including gradient diminishing[60] and weak temporal correlation structure in the neural predictive models.[61][62] 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 for speech recognition.[53][54][63][64]

2010s

By early 2010s *speech* recognition, also called voice recognition[65][66][67] was clearly differentiated from *speaker* recognition, and speaker independence was considered a major breakthrough. Until then, systems required a "training" period. A 1987 ad for a doll had carried the tagline "Finally, the doll that understands you." – despite the fact that it was described as "which children could train to respond to their voice".[12]

In 2017, Microsoft researchers reached a historical human parity milestone of transcribing conversational telephony speech on the widely benchmarked Switchboard task. Multiple deep learning models were used to optimize speech recognition accuracy. The speech recognition word error rate was reported to be as low as 4 professional human transcribers working together on the same benchmark, which was funded by IBM Watson speech team on the same task.[68]

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 that 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 that phonemes with different left and right context would have different realizations as HMM states); it would use cepstral normalization to normalize for a 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[69] (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.[70]

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,[71] phoneme classification through multi-objective evolutionary algorithms,[72] isolated word recognition,[73] audiovisual speech recognition, audiovisual speaker recognition and speaker adaptation.

Neural networks make fewer explicit assumptions about feature statistical properties than HMMs and have several qualities making them more 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. However, in spite of their effectiveness in classifying short-time units such as individual phonemes and isolated words,[74] early neural networks were rarely successful for continuous recognition tasks because of their limited ability to model temporal dependencies.

One approach to this limitation was to use neural networks as a pre-processing, feature transformation or dimensionality reduction,[75] step prior to HMM based recognition. However, more recently, LSTM and related recurrent neural networks (RNNs),[39][43][76][77] Time Delay Neural Networks(TDNN's),[78] and transformers[48][49][50] have demonstrated improved performance in this area.

Deep feedforward and recurrent neural networks

Deep neural networks and denoising autoencoders[79] are also under investigation. A deep feedforward neural network (DNN) is an artificial neural network with multiple hidden layers of units between the input and output layers.[53] 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.[80]

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.[81][82] [83] 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.[84] See also the related background of automatic speech recognition and the impact of various machine learning paradigms, notably including deep learning, in recent overview articles.[85][86]

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,[87] 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.[88]

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.[89] 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 at 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.[90] 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.[91] In 2016, University of Oxford presented LipNet,[92] the first end-to-end sentence-level lipreading model, using spatiotemporal convolutions coupled with an RNN-CTC architecture, surpassing human-level performance in a restricted grammar dataset.[93] A large-scale CNN-RNN-CTC architecture was presented in 2018 by Google DeepMind achieving 6 times better performance than human experts.[94] In 2019, Nvidia launched two CNN-CTC ASR models, Jasper and QuarzNet, with an overall performance WER of 3% [95][96]. Similar to other deep learning applications, transfer learning and domain adaptation are important strategies for reusing and extending the capabilities of deep learning models, particularly due to the high costs of training models from scratch, and the small size of available corpus in many languages and/or specific domains[97][98][99].

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 Bahdanau et al. of the University of Montreal in 2016.[100][101] 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 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).[102] 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;[103] University of Oxford and Google DeepMind extended LAS to "Watch, Listen, Attend and Spell" (WLAS) to handle lip reading surpassing human-level performance.[104]

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 signaled 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.

Education

Automatic pronunciation assessment is the use of speech recognition to verify the correctness of pronounced speech,[105] as distinguished from manual assessment by an instructor or proctor.[106] Also called speech verification, pronunciation evaluation, and pronunciation scoring, the main application of this technology is computer-aided pronunciation teaching (CAPT) when combined with computer-aided instruction for computer-assisted language learning (CALL), speech remediation, or accent reduction. Pronunciation assessment does not determine unknown speech (as in dictation or automatic transcription) but instead, knowing the expected word(s) in advance, it attempts to verify the correctness of the learner's pronunciation and ideally their intelligibility to listeners,[107][108] sometimes along with often inconsequential prosody such as intonation, pitch, tempo, rhythm, and stress.[109] Pronunciation assessment is also used in reading tutoring, for example in products such as Microsoft Teams[110] and from Amira Learning.[111] Automatic pronunciation assessment can also be used to help diagnose and treat speech disorders such as apraxia.[112]

Assessing authentic listener intelligibility is essential for avoiding inaccuracies from accent bias, especially in high-stakes assessments;[113][114][115] from words with multiple correct pronunciations;[116] and from phoneme coding errors in machine-readable pronunciation dictionaries.[117] In 2022, researchers found that some newer speech to text systems, based on end-to-end reinforcement learning to map audio signals directly into words, produce word and phrase confidence scores very closely correlated with genuine listener intelligibility.[118] In the Common European Framework of Reference for Languages (CEFR) assessment criteria for "overall phonological control", intelligibility outweighs formally correct pronunciation at all levels.[119]

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.

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.[120]

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,[121] 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.[122]

Speaker-independent systems are also being developed and are under test for the F-35 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%.[123]

Helicopters

The problems of achieving high recognition accuracy under stress and noise are particularly relevant in the helicopter environment as well as in 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 a 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. In telephony systems, ASR is now being predominantly used in contact centers by integrating it with IVR systems. Despite the high level of integration with word processing in general personal computing, in the field of document production, ASR 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.

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.[124]

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.[125]

Students who are physically disabled have a 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 enjoy searching the Internet or using a computer at home without having to physically operate a mouse and keyboard.[125]

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.[126] Also, see Learning disability.

The 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.

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.[127][128] 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.[129] 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.[130]

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 from 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.[131]

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[132]
- Automatic subtitling with speech recognition
- Automatic emotion recognition[133]
- Automatic shot listing in audiovisual production
- Automatic translation
- eDiscovery (Legal discovery)
- Hands-free computing: Speech recognition computer user interface
- Home automation
- Interactive voice response
- Mobile telephony, including mobile email
- Multimodal interaction[64]
- Real Time Captioning[134]
- Robotics
- Security, including usage with other biometric scanners for multi-factor authentication[135]
- Speech to text (transcription of speech into text, real time 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.[136][137] 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:[138]

- 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, the 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 confusing letters:

  e.g. the 26 letters of the English alphabet are difficult to discriminate because they are confusing words (most notoriously, the E-set: "B, C, D, E, G, P, T, V, Z — when "Z" is pronounced "zee" rather than "zed" depending on the English region); an 8% error rate is considered good for this vocabulary.[139]

- 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 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.
- Adverse conditions – Environmental noise (e.g. Noise in a car or a factory). Acoustical distortions (e.g. echoes, room acoustics)

Speech recognition is a multi-leveled 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 a lower level;

- This hierarchy of constraints is 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 into 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 the lower level, and going to lower levels, even more, we create more basic and shorter and simpler sounds. At 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 sounds on upper level, a new set of more deterministic rules should predict what the 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 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 two descriptions: amplitude (how strong is it), and frequency (how often it vibrates per second). Accuracy can be computed with the help of word error rate (WER). Word error rate can be calculated by aligning the recognized word and referenced word using dynamic string alignment. The problem may occur while computing the word error rate due to the difference between the sequence lengths of the recognized word and referenced word.

The formula to compute the word error rate (WER) is:

$$WER = \frac{(s+d+i)}{n}$$

where s is the number of substitutions, d is the number of deletions, i is the number of insertions, and n is the number of word references.

While computing, the word recognition rate (WRR) is used. The formula is:

$$WRR = 1 - WER = \frac{(n-s-d-i)}{n} = \frac{h-i}{n}$$

where h is the number of correctly recognized words:

$$h = n - (s+d).$$

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 can cause devices in homes and offices to start listening for input inappropriately, or possibly take an unwanted action.[140] Voice-controlled devices are also accessible to visitors to the building, or even those outside the building if they can be heard inside. Attackers may be able to gain access to personal information, like calendar, address book contents, private messages, and documents. They may also be able to impersonate the user to send messages or make online purchases.

Two attacks have been demonstrated that use artificial sounds. One transmits ultrasound and attempt to send commands without nearby people noticing.[141] The other adds small, inaudible distortions to other speech or music that are specially crafted to confuse the specific speech recognition system into recognizing music as speech, or to make what sounds like one command to a human sound like a different command to the system.[142]

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., "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 updated textbook *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 comprehensive textbook, "Fundamentals of Speaker Recognition" is an in depth source for up to date details on the theory and practice.[143] 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 Microsoft researchers D. Yu and L. Deng and 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.[84] 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.[80]

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.[144] In 2017 Mozilla launched the open source project called Common Voice[145] to gather big database of voices that would help build free speech recognition project DeepSpeech (available free at GitHub),[146] using Google's open source platform TensorFlow.[147] When Mozilla redirected funding away from the project in 2020, it was forked by its original developers as Coqui STT[148] using the same open-source license.[149][150]

Google Gboard supports speech recognition on all Android applications. It can be activated through the microphone icon.[151]

The commercial cloud based speech recognition APIs are broadly available.

For more software resources, see List of speech recognition software.

See also

- AI effect
- ALPAC
- Applications of artificial intelligence
- Articulatory speech recognition
- Audio mining
- Audio-visual speech recognition
- Automatic Language Translator
- Automotive head unit
- Braina
- Cache language model
- Dragon NaturallySpeaking
- Fluency Voice Technology
- Google Voice Search
- IBM ViaVoice
- Keyword spotting
- Kinect
- Mondegreen
- Multimedia information retrieval

- Origin of speech
- Phonetic search technology
- Speaker diarisation
- Speaker recognition
- Speech analytics
- Speech interface guideline
- Speech recognition software for Linux
- Speech synthesis
- Speech verification
- Subtitle (captioning)
- VoiceXML
- VoxForge
- Windows Speech Recognition

Lists

- List of speech recognition software
- List of emerging technologies
- Outline of artificial intelligence
- Timeline of speech and voice recognition

References

1. "Speaker Independent Connected Speech Recognition- Fifth Generation Computer Corporation" (http://www.fifthgen.com/speaker-independent-connected-s-r.htm). Fifthgen.com. Archived (https://web.archive.org/web/20131111101228/http://www.fifthgen.com/speaker-independent-connected-s-r.htm) from the original on 11 November 2013. Retrieved 15 June 2013.
2. P. Nguyen (2010). "Automatic classification of speaker characteristics". *International Conference on Communications and Electronics 2010*. pp. 147–152. doi:10.1109/ICCE.2010.5670700 (https://doi.org/10.1109%2FICCE.2010.5670700). ISBN 978-1-4244-7055-6. S2CID 13482115 (https://api.semanticscholar.org/CorpusID:13482115).
3. "British English definition of voice recognition" (http://www.macmillandictionary.com/dictionary/british/voice-recognition). Macmillan Publishers Limited. Archived (https://web.archive.org/web/20110916050430/http://www.macmillandictionary.com/dictionary/british/voice-recognition) from the original on 16 September 2011. Retrieved 21 February 2012.
4. "voice recognition, definition of" (http://www.businessdictionary.com/definition/voice-recognition.html). WebFinance, Inc. Archived (https://web.archive.org/web/20111203144647/http://www.businessdictionary.com/definition/voice-recognition.html) from the original on 3 December 2011. Retrieved 21 February 2012.
5. "The Mailbag LG #114" (http://linuxgazette.net/114/lg_mail.html#mailbag.3). Linuxgazette.net. Archived (https://web.archive.org/web/20130219032501/http://linuxgazette.net/114/lg_mail.html#mailbag.3) from the original on 19 February 2013. Retrieved 15 June 2013.
6. Sarangi, Susanta; Sahidullah, Md; Saha, Goutam (September 2020). "Optimization of data-driven filterbank for automatic speaker verification". *Digital Signal Processing*. **104**: 102795. arXiv:2007.10729 (https://arxiv.org/abs/2007.10729). Bibcode:2020DSP...10402795S (https://ui.adsabs.harvard.edu/abs/2020DSP...10402795S). doi:10.1016/j.dsp.2020.102795 (https://doi.org/10.1016%2Fj.dsp.2020.102795). S2CID 220665533 (https://api.semanticscholar.org/CorpusID:220665533).

7. Reynolds, Douglas; Rose, Richard (January 1995). "Robust text-independent speaker identification using Gaussian mixture speaker models" (http://www.cs.toronto.edu/~frank/csc401/readings/ReynoldsRose.pdf) (PDF). *IEEE Transactions on Speech and Audio Processing*. **3** (1): 72–83. doi:10.1109/89.365379 (https://doi.org/10.1109%2F89.365379). ISSN 1063-6676 (https://search.worldcat.org/issn/1063-6676). OCLC 26108901 (https://search.worldcat.org/oclc/26108901). S2CID 7319345 (https://api.semanticscholar.org/CorpusID:7319345). Archived (https://web.archive.org/web/20140308001101/http://www.cs.toronto.edu/~frank/csc401/readings/ReynoldsRose.pdf) (PDF) from the original on 8 March 2014. Retrieved 21 February 2014.
8. "Speaker Identification (WhisperID)" (http://research.microsoft.com/en-us/projects/whisperid/). *Microsoft Research*. Microsoft. Archived (https://web.archive.org/web/20140225190956/http://research.microsoft.com/en-us/projects/whisperid/) from the original on 25 February 2014. Retrieved 21 February 2014. "When you speak to someone, they don't just recognize what you say: they recognize who you are. WhisperID will let computers do that, too, figuring out who you are by the way you sound."
9. "Obituaries: Stephen Balashek" (https://obits.nj.com/obituaries/starledger/obituary.aspx?page=lifestory&pid=158702138). *The Star-Ledger*. 22 July 2012. Archived (https://web.archive.org/web/20190404231352/https://obits.nj.com/obituaries/starledger/obituary.aspx?page=lifestory&pid=158702138) from the original on 4 April 2019. Retrieved 9 September 2024.
10. "IBM-Shoebox-front.jpg" (https://cdn57.androidauthority.net/wp-content/uploads/2012/04/IBM-Shoebox-front.jpg). androidauthority.net. Archived (https://web.archive.org/web/20180809153221/https://cdn57.androidauthority.net/wp-content/uploads/2012/04/IBM-Shoebox-front.jpg) from the original on 9 August 2018. Retrieved 4 April 2019.
11. Juang, B. H.; Rabiner, Lawrence R. "Automatic speech recognition–a brief history of the technology development" (http://www.ece.ucsb.edu/faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf) (PDF). p. 6. Archived (https://web.archive.org/web/20140817193243/http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf) (PDF) from the original on 17 August 2014. Retrieved 17 January 2015.
12. Melanie Pinola (2 November 2011). "Speech Recognition Through the Decades: How We Ended Up With Siri" (https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html). *PC World*. Archived (https://web.archive.org/web/20181103105727/https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html) from the original on 3 November 2018. Retrieved 22 October 2018.
13. Gray, Robert M. (2010). "A History of Realtime Digital Speech on Packet Networks: Part II of Linear Predictive Coding and the Internet Protocol" (https://ee.stanford.edu/~gray/lpcip.pdf) (PDF). *Found. Trends Signal Process*. **3** (4): 203–303. doi:10.1561/2000000036 (https://doi.org/10.1561%2F2000000036). ISSN 1932-8346 (https://search.worldcat.org/issn/1932-8346). Archived (https://ghostarchive.org/archive/20221009/https://ee.stanford.edu/~gray/lpcip.pdf) (PDF) from the original on 9 October 2022. Retrieved 9 September 2024.
14. John R. Pierce (1969). "Whither speech recognition?". *Journal of the Acoustical Society of America*. **46** (48): 1049–1051. Bibcode:1969ASAJ...46.1049P (https://ui.adsabs.harvard.edu/abs/1969ASAJ...46.1049P). doi:10.1121/1.1911801 (https://doi.org/10.1121%2F1.1911801).
15. Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (2008). *Springer Handbook of Speech Processing*. Springer Science & Business Media. ISBN 978-3540491255.

16. John Makhoul. "ISCA Medalist: For leadership and extensive contributions to speech and language processing" (https://www.superlectures.com/interspeech2016/isca-medalist-for-leadership-and-extensive-contributions-to-speech-and-language-processing). Archived (https://web.archive.org/web/20180124071005/https://www.superlectures.com/interspeech2016/isca-medalist-for-leadership-and-extensive-contributions-to-speech-and-language-processing) from the original on 24 January 2018. Retrieved 23 January 2018.
17. Blechman, R. O.; Blechman, Nicholas (23 June 2008). "Hello, Hal" (https://www.newyorker.com/magazine/2008/06/23/hello-hal). *The New Yorker*. Archived (https://web.archive.org/web/20150120042048/http://www.newyorker.com/magazine/2008/06/23/hello-hal) from the original on 20 January 2015. Retrieved 17 January 2015.
18. Klatt, Dennis H. (1977). "Review of the ARPA speech understanding project". *The Journal of the Acoustical Society of America*. **62** (6): 1345–1366. Bibcode:1977ASAJ...62.1345K (https://ui.adsabs.harvard.edu/abs/1977ASAJ...62.1345K). doi:10.1121/1.381666 (https://doi.org/10.1121%2F1.381666).
19. Rabiner (1984). "The Acoustics, Speech, and Signal Processing Society. A Historical Perspective" (http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/216_historical%20perspective.pdf) (PDF). Archived (https://web.archive.org/web/20170809113828/http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/216_historical%20perspective.pdf) (PDF) from the original on 9 August 2017. Retrieved 23 January 2018.
20. "First-Hand:The Hidden Markov Model – Engineering and Technology History Wiki" (http://ethw.org/First-Hand:The_Hidden_Markov_Model). *ethw.org*. 12 January 2015. Archived (https://web.archive.org/web/20180403191314/http://ethw.org/First-Hand:The_Hidden_Markov_Model) from the original on 3 April 2018. Retrieved 1 May 2018.
21. "James Baker interview" (http://www.sarasinstitute.org/Audio/JimBaker(2006).mp3). Archived (https://web.archive.org/web/20170828105222/http://www.sarasinstitute.org/Audio/JimBaker(2006).mp3) from the original on 28 August 2017. Retrieved 9 February 2017.
22. "Pioneering Speech Recognition" (https://web.archive.org/web/20150219080748/http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/). 7 March 2012. Archived from the original (http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/) on 19 February 2015. Retrieved 18 January 2015.
23. Huang, Xuedong; Baker, James; Reddy, Raj (January 2014). "A historical perspective of speech recognition" (https://web.archive.org/web/20231208161616/https://dl.acm.org/doi/fullHtml/10.1145/2500887). *Communications of the ACM*. **57** (1): 94–103. doi:10.1145/2500887 (https://doi.org/10.1145%2F2500887). ISSN 0001-0782 (https://search.worldcat.org/issn/0001-0782). S2CID 6175701 (https://api.semanticscholar.org/CorpusID:6175701). Archived from the original (https://dl.acm.org/doi/fullHtml/10.1145/2500887) on 8 December 2023.
24. Juang, B. H.; Rabiner, Lawrence R. Automatic speech recognition–a brief history of the technology development (http://www.ece.ucsb.edu/faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf) (PDF) (Report). p. 10. Archived (https://web.archive.org/web/20140817193243/http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf) (PDF) from the original on 17 August 2014. Retrieved 17 January 2015.
25. Li, Xiaochang (1 July 2023). " "There's No Data Like More Data": Automatic Speech Recognition and the Making of Algorithmic Culture" (https://www.journals.uchicago.edu/doi/10.1086/725132). *Osiris*. **38**: 165–182. doi:10.1086/725132 (https://doi.org/10.1086%2F725132). ISSN 0369-7827 (https://search.worldcat.org/issn/0369-7827). S2CID 259502346 (https://api.semanticscholar.org/CorpusID:259502346).

26. "History of Speech Recognition" (https://web.archive.org/web/20150813223326/http://dragon-medical-transcription.com/history_speech_recognition.html). *Dragon Medical Transcription*. Archived from the original (http://www.dragon-medical-transcription.com/history_speech_recognition.html) on 13 August 2015. Retrieved 17 January 2015.
27. Billi, Roberto; Canavesio, Franco; Ciaramella, Alberto; Nebbia, Luciano (1 November 1995). "Interactive voice technology at work: The CSELT experience" (https://www.sciencedirect.com/science/article/abs/pii/016763939500030R). *Speech Communication*. **17** (3): 263–271. doi:10.1016/0167-6393(95)00030-R (https://doi.org/10.1016%2F0167-6393%2895%2900030-R).
28. Xuedong Huang; James Baker; Raj Reddy (January 2014). "A Historical Perspective of Speech Recognition" (http://cacm.acm.org/magazines/2014/1/170863-a-historical-perspective-of-speech-recognition/fulltext#R5). Communications of the ACM. Archived (http://archive.wikiwix.com/cache/20150120074239/http://cacm.acm.org/magazines/2014/1/170863-a-historical-perspective-of-speech-recognition/fulltext#R5) from the original on 20 January 2015. Retrieved 20 January 2015.
29. Kevin McKean (8 April 1980). "When Cole talks, computers listen" (https://news.google.com/newspapers?nid=1798&dat=19800408&id=xgsdAAAAIBAJ&pg=6057,1141823). Sarasota Journal. AP. Retrieved 23 November 2015.
30. "ACT/Apricot - Apricot history" (http://actapricot.org/history/apricot_review_1.html). *actapricot.org*. Archived (https://web.archive.org/web/20161221091131/http://actapricot.org/history/apricot_review_1.html) from the original on 21 December 2016. Retrieved 2 February 2016.
31. Melanie Pinola (2 November 2011). "Speech Recognition Through the Decades: How We Ended Up With Siri" (http://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html?page=2). *PC World*. Archived (https://web.archive.org/web/20170113074944/http://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html?page=2) from the original on 13 January 2017. Retrieved 28 July 2017.
32. "Ray Kurzweil biography" (http://www.kurzweilai.net/ray-kurzweil-bio). KurzweilAINetwork. Archived (https://web.archive.org/web/20140205002828/http://www.kurzweilai.net/ray-kurzweil-bio) from the original on 5 February 2014. Retrieved 25 September 2014.
33. Juang, B.H.; Rabiner, Lawrence. Automatic Speech Recognition – A Brief History of the Technology Development (http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf) (PDF) (Report). Archived (https://web.archive.org/web/20170809211311/http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf) (PDF) from the original on 9 August 2017. Retrieved 28 July 2017.
34. "Nuance Exec on iPhone 4S, Siri, and the Future of Speech" (http://techpinions.com/nuance-exec-on-iphone-4s-siri-and-the-future-of-speech/3307). Tech.pinions. 10 October 2011. Archived (https://web.archive.org/web/20111119211021/http://techpinions.com/nuance-exec-on-iphone-4s-siri-and-the-future-of-speech/3307) from the original on 19 November 2011. Retrieved 23 November 2011.
35. "Switchboard-1 Release 2" (https://catalog.ldc.upenn.edu/LDC97S62). Archived (https://web.archive.org/web/20170711061225/https://catalog.ldc.upenn.edu/LDC97S62) from the original on 11 July 2017. Retrieved 26 July 2017.
36. Jason Kincaid (13 February 2011). "The Power of Voice: A Conversation With The Head Of Google's Speech Technology" (https://techcrunch.com/2011/02/13/the-power-of-voice-a-conversation-with-the-head-of-googles-speech-technology/). *Tech Crunch*. Archived (https://web.archive.org/web/20150721034447/http://techcrunch.com/2011/02/13/the-power-of-voice-a-conversation-with-the-head-of-googles-speech-technology/) from the original on 21 July 2015. Retrieved 21 July 2015.
37. Froomkin, Dan (5 May 2015). "THE COMPUTERS ARE LISTENING" (https://firstlook.org/theintercept/2015/05/05/nsa-speech-recognition-snowden-searchable-text/). *The Intercept*. Archived (https://web.archive.org/web/20150627185007/https://firstlook.org/theintercept/2015/05/05/nsa-speech-recognition-snowden-searchable-text/) from the original on 27 June 2015. Retrieved 20 June 2015.

38. Herve Bourlard and Nelson Morgan, Connectionist Speech Recognition: A Hybrid Approach, The Kluwer International Series in Engineering and Computer Science; v. 247, Boston: Kluwer Academic Publishers, 1994.
39. Sepp Hochreiter; J. Schmidhuber (1997). "Long Short-Term Memory". *Neural Computation*. **9** (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735 (https://doi.org/10.1162%2Fneco.1997.9.8.1735). PMID 9377276 (https://pubmed.ncbi.nlm.nih.gov/9377276). S2CID 1915014 (https://api.semanticscholar.org/CorpusID:1915014).
40. Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". *Neural Networks*. **61**: 85–117. arXiv:1404.7828 (https://arxiv.org/abs/1404.7828). doi:10.1016/j.neunet.2014.09.003 (https://doi.org/10.1016%2Fj.neunet.2014.09.003). PMID 25462637 (https://pubmed.ncbi.nlm.nih.gov/25462637). S2CID 11715509 (https://api.semanticscholar.org/CorpusID:11715509).
41. Alex Graves, Santiago Fernandez, Faustino Gomez, and Jürgen Schmidhuber (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets (https://mediatum.ub.tum.de/doc/1292048/file.pdf) Archived (https://web.archive.org/web/20240909053409/https://mediatum.ub.tum.de/doc/1292048/file.pdf) 9 September 2024 at the Wayback Machine. Proceedings of ICML'06, pp. 369–376.
42. Santiago Fernandez, Alex Graves, and Jürgen Schmidhuber (2007). An application of recurrent neural networks to discriminative keyword spotting (https://www6.in.tum.de/pub/Main/Publications/Fernandez2007b.pdf). Proceedings of ICANN (2), pp. 220–229.
43. Haşim Sak, Andrew Senior, Kanishka Rao, Françoise Beaufays and Johan Schalkwyk (September 2015): ""Google voice search: faster and more accurate" (https://web.archive.org/web/20160309191532/http://googleresearch.blogspot.ch/2015/09/google-voice-search-faster-and-more.html). Archived from the original (http://googleresearch.blogspot.ch/2015/09/google-voice-search-faster-and-more.html) on 9 March 2016. Retrieved 5 April 2016.."
44. Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob; Houlsby, Neil (3 June 2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". arXiv:2010.11929 (https://arxiv.org/abs/2010.11929) [cs.CV (https://arxiv.org/archive/cs.CV)].
45. Wu, Haiping; Xiao, Bin; Codella, Noel; Liu, Mengchen; Dai, Xiyang; Yuan, Lu; Zhang, Lei (29 March 2021). "CvT: Introducing Convolutions to Vision Transformers". arXiv:2103.15808 (https://arxiv.org/abs/2103.15808) [cs.CV (https://arxiv.org/archive/cs.CV)].
46. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need" (https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html). *Advances in Neural Information Processing Systems*. **30**. Curran Associates. Archived (https://web.archive.org/web/20240909053411/https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html) from the original on 9 September 2024. Retrieved 9 September 2024.
47. Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (24 May 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805 (https://arxiv.org/abs/1810.04805) [cs.CL (https://arxiv.org/archive/cs.CL)].
48. Gong, Yuan; Chung, Yu-An; Glass, James (8 July 2021). "AST: Audio Spectrogram Transformer". arXiv:2104.01778 (https://arxiv.org/abs/2104.01778) [cs.SD (https://arxiv.org/archive/cs.SD)].
49. Ristea, Nicolae-Catalin; Ionescu, Radu Tudor; Khan, Fahad Shahbaz (20 June 2022). "SepTr: Separable Transformer for Audio Spectrogram Processing". arXiv:2203.09581 (https://arxiv.org/abs/2203.09581) [cs.CV (https://arxiv.org/archive/cs.CV)].

50. Lohrenz, Timo; Li, Zhengyang; Fingscheidt, Tim (14 July 2021). "Multi-Encoder Learning and Stream Fusion for Transformer-Based End-to-End Automatic Speech Recognition". arXiv:2104.00120 (https://arxiv.org/abs/2104.00120) [eess.AS (https://arxiv.org/archive/eess.AS)].
51. "Li Deng" (https://lidengsite.wordpress.com/). Li Deng Site. Archived (https://web.archive.org/web/20240909052323/https://lidengsite.wordpress.com/) from the original on 9 September 2024. Retrieved 9 September 2024.
52. NIPS Workshop: Deep Learning for Speech Recognition and Related Applications, Whistler, BC, Canada, Dec. 2009 (Organizers: Li Deng, Geoff Hinton, D. Yu).
53. Hinton, Geoffrey; Deng, Li; Yu, Dong; Dahl, George; Mohamed, Abdel-Rahman; Jaitly, Navdeep; Senior, Andrew; Vanhoucke, Vincent; Nguyen, Patrick; Sainath, Tara; Kingsbury, Brian (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups". *IEEE Signal Processing Magazine*. **29** (6): 82–97. Bibcode:2012ISPM...29...82H (https://ui.adsabs.harvard.edu/abs/2012ISPM...29...82H). doi:10.1109/MSP.2012.2205597 (https://doi.org/10.1109%2FMSP.2012.2205597). S2CID 206485943 (https://api.semanticscholar.org/CorpusID:206485943).
54. Deng, L.; Hinton, G.; Kingsbury, B. (2013). "New types of deep neural network learning for speech recognition and related applications: An overview". *2013 IEEE International Conference on Acoustics, Speech and Signal Processing: New types of deep neural network learning for speech recognition and related applications: An overview*. p. 8599. doi:10.1109/ICASSP.2013.6639344 (https://doi.org/10.1109%2FICASSP.2013.6639344). ISBN 978-1-4799-0356-6. S2CID 13953660 (https://api.semanticscholar.org/CorpusID:13953660).
55. Markoff, John (23 November 2012). "Scientists See Promise in Deep-Learning Programs" (https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html). *New York Times*. Archived (https://web.archive.org/web/20121130080314/http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html) from the original on 30 November 2012. Retrieved 20 January 2015.
56. Morgan, Bourlard, Renals, Cohen, Franco (1993) "Hybrid neural network/hidden Markov model systems for continuous speech recognition. ICASSP/IJPRAI"
57. T. Robinson (1992). "A real-time recurrent error propagation network word recognition system" (https://www.researchgate.net/publication/3532171). *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 617–620 vol.1. doi:10.1109/ICASSP.1992.225833 (https://doi.org/10.1109%2FICASSP.1992.225833). ISBN 0-7803-0532-9. S2CID 62446313 (https://api.semanticscholar.org/CorpusID:62446313).
58. Waibel, Hanazawa, Hinton, Shikano, Lang. (1989) "Phoneme recognition using time-delay neural networks (http://www.inf.ufrgs.br/~engel/data/media/file/cmp121/waibel89_TDNN.pdf) Archived (https://web.archive.org/web/20210225163001/http://www.inf.ufrgs.br/~engel/data/media/file/cmp121/waibel89_TDNN.pdf) 25 February 2021 at the Wayback Machine. IEEE Transactions on Acoustics, Speech, and Signal Processing."
59. Baker, J.; Li Deng; Glass, J.; Khudanpur, S.; Chin-Hui Lee; Morgan, N.; O'Shaughnessy, D. (2009). "Developments and Directions in Speech Recognition and Understanding, Part 1". *IEEE Signal Processing Magazine*. **26** (3): 75–80. Bibcode:2009ISPM...26...75B (https://ui.adsabs.harvard.edu/abs/2009ISPM...26...75B). doi:10.1109/MSP.2009.932166 (https://doi.org/10.1109%2FMSP.2009.932166). hdl:1721.1/51891 (https://hdl.handle.net/1721.1%2F51891). S2CID 357467 (https://api.semanticscholar.org/CorpusID:357467).

60. Sepp Hochreiter (1991), Untersuchungen zu dynamischen neuronalen Netzen (http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf) Archived (https://web.archive.org/web/20150306075401/http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf) 6 March 2015 at the Wayback Machine, Diploma thesis. Institut f. Informatik, Technische Univ. Munich. Advisor: J. Schmidhuber.
61. Bengio, Y. (1991). *Artificial Neural Networks and their Application to Speech/Sequence Recognition* (https://elibrary.ru/item.asp?id=5790854) (Ph.D. thesis). McGill University.
62. Deng, L.; Hassanein, K.; Elmasry, M. (1994). "Analysis of the correlation structure for a neural predictive model with application to speech recognition". *Neural Networks*. **7** (2): 331–339. doi:10.1016/0893-6080(94)90027-2 (https://doi.org/10.1016%2F0893-6080%2894%2990027-2).
63. Keynote talk: Recent Developments in Deep Neural Networks. ICASSP, 2013 (by Geoff Hinton).
64. Keynote talk: "Achievements and Challenges of Deep Learning: From Speech Analysis and Recognition To Language and Multimodal Processing (https://www.isca-speech.org/archive/interspeech_2014/i14_3505.html) Archived (https://web.archive.org/web/20210305043518/https://www.isca-speech.org/archive/interspeech_2014/i14_3505.html) 5 March 2021 at the Wayback Machine," Interspeech, September 2014 (by Li Deng).
65. "Improvements in voice recognition software increase" (https://web.archive.org/web/20181023080207/https://www.techrepublic.com/article/improvements-in-voice-recognition-software-increase-productivity/). *TechRepublic.com*. 27 August 2002. Archived from the original (https://www.techrepublic.com/article/improvements-in-voice-recognition-software-increase-productivity) on 23 October 2018. Retrieved 22 October 2018. "Maners said IBM has worked on advancing speech recognition ... or on the floor of a noisy trade show."
66. "Voice Recognition To Ease Travel Bookings: Business Travel News" (http://www.businesstravelnews.com/More-News/Voice-Recognition-To-Ease-Travel-Bookings). *BusinessTravelNews.com*. 3 March 1997. Archived (https://web.archive.org/web/20240909052252/https://www.businesstravelnews.com/More-News/Voice-Recognition-To-Ease-Travel-Bookings) from the original on 9 September 2024. Retrieved 9 September 2024. "The earliest applications of speech recognition software were dictation ... Four months ago, IBM introduced a 'continual dictation product' designed to ... debuted at the National Business Travel Association trade show in 1994."
67. Ellis Booker (14 March 1994). "Voice recognition enters the mainstream". *Computerworld*. p. 45. "Just a few years ago, speech recognition was limited to ..."
68. "Microsoft researchers achieve new conversational speech recognition milestone" (https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/). *Microsoft*. 21 August 2017. Archived (https://web.archive.org/web/20240909052234/https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/) from the original on 9 September 2024. Retrieved 9 September 2024.
69. Goel, Vaibhava; Byrne, William J. (2000). "Minimum Bayes-risk automatic speech recognition" (http://www.clsp.jhu.edu/people/vgoel/publications/CSAL.ps). *Computer Speech & Language*. **14** (2): 115–135. doi:10.1006/csla.2000.0138 (https://doi.org/10.1006%2Fcsla.2000.0138). S2CID 206561058 (https://api.semanticscholar.org/CorpusID:206561058). Archived (https://web.archive.org/web/20110725225846/http://www.clsp.jhu.edu/people/vgoel/publications/CSAL.ps) from the original on 25 July 2011. Retrieved 28 March 2011.

70. Mohri, M. (2002). "Edit-Distance of Weighted Automata: General Definitions and Algorithms" (http://www.cs.nyu.edu/~mohri/pub/edit.pdf) (PDF). *International Journal of Foundations of Computer Science*. **14** (6): 957–982. doi:10.1142/S0129054103002114 (https://doi.org/10.1142%2FS0129054103002114). Archived (https://web.archive.org/web/20120318032640/http://www.cs.nyu.edu/~mohri/pub/edit.pdf) (PDF) from the original on 18 March 2012. Retrieved 28 March 2011.
71. Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K. J. (1989). "Phoneme recognition using time-delay neural networks". *IEEE Transactions on Acoustics, Speech, and Signal Processing*. **37** (3): 328–339. doi:10.1109/29.21701 (https://doi.org/10.1109%2F29.21701). hdl:10338.dmlcz/135496 (https://hdl.handle.net/10338.dmlcz%2F135496). S2CID 9563026 (https://api.semanticscholar.org/CorpusID:9563026).
72. Bird, Jordan J.; Wanner, Elizabeth; Ekárt, Anikó; Faria, Diego R. (2020). "Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms" (https://publications.aston.ac.uk/id/eprint/41416/1/Speech_Recog_ESWA_2_.pdf) (PDF). *Expert Systems with Applications*. **153**. Elsevier BV: 113402. doi:10.1016/j.eswa.2020.113402 (https://doi.org/10.1016%2Fj.eswa.2020.113402). ISSN 0957-4174 (https://search.worldcat.org/issn/0957-4174). S2CID 216472225 (https://api.semanticscholar.org/CorpusID:216472225). Archived (https://web.archive.org/web/20240909053419/https://publications.aston.ac.uk/id/eprint/41416/1/Speech_Recog_ESWA_2_.pdf) (PDF) from the original on 9 September 2024. Retrieved 9 September 2024.
73. Wu, J.; Chan, C. (1993). "Isolated Word Recognition by Neural Network Models with Cross-Correlation Coefficients for Speech Dynamics". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **15** (11): 1174–1185. doi:10.1109/34.244678 (https://doi.org/10.1109%2F34.244678).
74. S. A. Zahorian, A. M. Zimmer, and F. Meng, (2002) "Vowel Classification for Computer based Visual Feedback for Speech Training for the Hearing Impaired (https://www.researchgate.net/profile/Stephen_Zahorian/publication/221480228_Vowel_classification_for_computer-based_visual_feedback_for_speech_training_for_the_hearing_impaired/links/00b7d525d25f51c585000000.pdf)," in ICSLP 2002
75. Hu, Hongbing; Zahorian, Stephen A. (2010). "Dimensionality Reduction Methods for HMM Phonetic Recognition" (http://bingweb.binghamton.edu/~hhu1/paper/Hu2010Dimensionality.pdf) (PDF). *ICASSP 2010*. Archived (http://archive.wikiwix.com/cache/20120706063756/http://bingweb.binghamton.edu/~hhu1/paper/Hu2010Dimensionality.pdf) (PDF) from the original on 6 July 2012.
76. Fernandez, Santiago; Graves, Alex; Schmidhuber, Jürgen (2007). "Sequence labelling in structured domains with hierarchical recurrent neural networks" (http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-124.pdf) (PDF). *Proceedings of IJCAI*. Archived (https://web.archive.org/web/20170815003130/http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-124.pdf) (PDF) from the original on 15 August 2017.
77. Graves, Alex; Mohamed, Abdel-rahman; Hinton, Geoffrey (2013). "Speech recognition with deep recurrent neural networks". arXiv:1303.5778 (https://arxiv.org/abs/1303.5778) [cs.NE (https://arxiv.org/archive/cs.NE)]. ICASSP 2013.
78. Waibel, Alex (1989). "Modular Construction of Time-Delay Neural Networks for Speech Recognition" (http://isl.anthropomatik.kit.edu/cmu-kit/Modular_Construction_of_Time-Delay_Neural_Networks_for_Speech_Recognition.pdf) (PDF). *Neural Computation*. **1** (1): 39–46. doi:10.1162/neco.1989.1.1.39 (https://doi.org/10.1162%2Fneco.1989.1.1.39). S2CID 236321 (https://api.semanticscholar.org/CorpusID:236321). Archived (https://web.archive.org/web/20160629180846/http://isl.anthropomatik.kit.edu/cmu-kit/Modular_Construction_of_Time-Delay_Neural_Networks_for_Speech_Recognition.pdf) (PDF) from the original on 29 June 2016.
79. Maas, Andrew L.; Le, Quoc V.; O'Neil, Tyler M.; Vinyals, Oriol; Nguyen, Patrick; Ng, Andrew Y. (2012). "Recurrent Neural Networks for Noise Reduction in Robust ASR". *Proceedings of Interspeech 2012*.

80. Deng, Li; Yu, Dong (2014). "Deep Learning: Methods and Applications" (http://research.microsoft.com/pubs/209355/DeepLearning-NowPublishing-Vol7-SIG-039.pdf) (PDF). *Foundations and Trends in Signal Processing*. **7** (3–4): 197–387. CiteSeerX 10.1.1.691.3679 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.691.3679). doi:10.1561/2000000039 (https://doi.org/10.1561%2F2000000039). Archived (https://web.archive.org/web/20141022161017/http://research.microsoft.com/pubs/209355/DeepLearning-NowPublishing-Vol7-SIG-039.pdf) (PDF) from the original on 22 October 2014.
81. Yu, D.; Deng, L.; Dahl, G. (2010). "Roles of Pre-Training and Fine-Tuning in Context-Dependent DBN-HMMs for Real-World Speech Recognition" (https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dbn4asr-nips2010.pdf) (PDF). *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.
82. Dahl, George E.; Yu, Dong; Deng, Li; Acero, Alex (2012). "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition". *IEEE Transactions on Audio, Speech, and Language Processing*. **20** (1): 30–42. doi:10.1109/TASL.2011.2134090 (https://doi.org/10.1109%2FTASL.2011.2134090). S2CID 14862572 (https://api.semanticscholar.org/CorpusID:14862572).
83. Deng L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F. et al. Recent Advances in Deep Learning for Speech Research at Microsoft (https://pdfs.semanticscholar.org/6bdc/cfe195bc49d218acc5be750aa49e41f408e4.pdf) Archived (https://web.archive.org/web/20240909052236/https://pdfs.semanticscholar.org/6bdc/cfe195bc49d218acc5be750aa49e41f408e4.pdf) 9 September 2024 at the Wayback Machine. ICASSP, 2013.
84. Yu, D.; Deng, L. (2014). "Automatic Speech Recognition: A Deep Learning Approach (Publisher: Springer)". {{cite journal}}: Cite journal requires |journal= (help)
85. Deng, L.; Li, Xiao (2013). "Machine Learning Paradigms for Speech Recognition: An Overview" (http://cvsp.cs.ntua.gr/courses/patrec/slides_material2018/slides-2018/DengLi_MLParadigms-SpeechRecogn-AnOverview_TALSP13.pdf) (PDF). *IEEE Transactions on Audio, Speech, and Language Processing*. **21** (5): 1060–1089. doi:10.1109/TASL.2013.2244083 (https://doi.org/10.1109%2FTASL.2013.2244083). S2CID 16585863 (https://api.semanticscholar.org/CorpusID:16585863). Archived (https://web.archive.org/web/20240909052239/http://cvsp.cs.ntua.gr/courses/patrec/slides_material2018/slides-2018/DengLi_MLParadigms-SpeechRecogn-AnOverview_TALSP13.pdf) (PDF) from the original on 9 September 2024. Retrieved 9 September 2024.
86. Schmidhuber, Jürgen (2015). "Deep Learning" (https://doi.org/10.4249%2Fscholarpedia.32832). *Scholarpedia*. **10** (11): 32832. Bibcode:2015SchpJ..1032832S (https://ui.adsabs.harvard.edu/abs/2015SchpJ..1032832S). doi:10.4249/scholarpedia.32832 (https://doi.org/10.4249%2Fscholarpedia.32832).
87. L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton (2010) Binary Coding of Speech Spectrograms Using a Deep Auto-encoder (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.185.1908&rep=rep1&type=pdf). Interspeech.
88. Tüske, Zoltán; Golik, Pavel; Schlüter, Ralf; Ney, Hermann (2014). "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR" (https://www-i6.informatik.rwth-aachen.de/publications/download/937/T%7Bu%7DskeZolt%7Ba%7DnGolikPavelSchl%7Bu%7DterRalfNeyHermann--AcousticModelingwithDeepNeuralNetworksUsingRawTimeSignalfor%7BLVCSR%7D--2014.pdf) (PDF). *Interspeech 2014*. Archived (https://web.archive.org/web/20161221174753/https://www-i6.informatik.rwth-aachen.de/publications/download/937/T%7Bu%7DskeZolt%7Ba%7DnGolikPavelSchl%7Bu%7DterRalfNeyHermann--AcousticModelingwithDeepNeuralNetworksUsingRawTimeSignalfor%7BLVCSR%7D--2014.pdf) (PDF) from the original on 21 December 2016.
89. Jurafsky, Daniel (2016). *Speech and Language Processing*.

90. Graves, Alex (2014). "Towards End-to-End Speech Recognition with Recurrent Neural Networks" (https://web.archive.org/web/20170110184531/http://jmlr.org/proceedings/papers/v32/graves14.pdf) (PDF). *ICML*. Archived from the original (http://www.jmlr.org/proceedings/papers/v32/graves14.pdf) (PDF) on 10 January 2017. Retrieved 22 July 2019.
91. Amodei, Dario (2016). "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". arXiv:1512.02595 (https://arxiv.org/abs/1512.02595) [cs.CL (https://arxiv.org/archive/cs.CL)].
92. "LipNet: How easy do you think lipreading is?" (https://www.youtube.com/watch?v=fa5QGremQf8). *YouTube*. 4 November 2016. Archived (https://web.archive.org/web/20170427104009/https://www.youtube.com/watch?v=fa5QGremQf8) from the original on 27 April 2017. Retrieved 5 May 2017.
93. Assael, Yannis; Shillingford, Brendan; Whiteson, Shimon; de Freitas, Nando (5 November 2016). "LipNet: End-to-End Sentence-level Lipreading". arXiv:1611.01599 (https://arxiv.org/abs/1611.01599) [cs.CV (https://arxiv.org/archive/cs.CV)].
94. Shillingford, Brendan; Assael, Yannis; Hoffman, Matthew W.; Paine, Thomas; Hughes, Cían; Prabhu, Utsav; Liao, Hank; Sak, Hasim; Rao, Kanishka (13 July 2018). "Large-Scale Visual Speech Recognition". arXiv:1807.05162 (https://arxiv.org/abs/1807.05162) [cs.CV (https://arxiv.org/archive/cs.CV)].
95. Li, Jason; Lavrukhin, Vitaly; Ginsburg, Boris; Leary, Ryan; Kuchaiev, Oleksii; Cohen, Jonathan M.; Nguyen, Huyen; Gadde, Ravi Teja (2019). "Jasper: An End-to-End Convolutional Neural Acoustic Model" (https://www.isca-archive.org/interspeech_2019/li19_interspeech.html). *Interspeech 2019*. pp. 71–75. arXiv:1904.03288 (https://arxiv.org/abs/1904.03288). doi:10.21437/Interspeech.2019-1819 (https://doi.org/10.21437%2FInterspeech.2019-1819).
96. Kriman, Samuel; Beliaev, Stanislav; Ginsburg, Boris; Huang, Jocelyn; Kuchaiev, Oleksii; Lavrukhin, Vitaly; Leary, Ryan; Li, Jason; Zhang, Yang (22 October 2019), *QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions* (https://arxiv.org/abs/1910.10261), arXiv:1910.10261 (https://arxiv.org/abs/1910.10261), retrieved 30 September 2024
97. Medeiros, Eduardo; Corado, Leonel; Rato, Luís; Quaresma, Paulo; Salgueiro, Pedro (May 2023). "Domain Adaptation Speech-to-Text for Low-Resource European Portuguese Using Deep Learning" (https://doi.org/10.3390%2Ffi15050159). *Future Internet*. **15** (5): 159. doi:10.3390/fi15050159 (https://doi.org/10.3390%2Ffi15050159). ISSN 1999-5903 (https://search.worldcat.org/issn/1999-5903).
98. Joshi, Raviraj; Singh, Anupam (May 2022). Malmasi, Shervin; Rokhlenko, Oleg; Ueffing, Nicola; Guy, Ido; Agichtein, Eugene; Kallumadi, Surya (eds.). "A Simple Baseline for Domain Adaptation in End to End ASR Systems Using Synthetic Data" (https://aclanthology.org/2022.ecnlp-1.28/). *Proceedings of the Fifth Workshop on E-Commerce and NLP (ECNLP 5)*. Dublin, Ireland: Association for Computational Linguistics: 244–249. doi:10.18653/v1/2022.ecnlp-1.28 (https://doi.org/10.18653%2Fv1%2F2022.ecnlp-1.28).
99. Sukhadia, Vrunda N.; Umesh, S. (9 January 2023). "Domain Adaptation of Low-Resource Target-Domain Models Using Well-Trained ASR Conformer Models" (https://ieeexplore.ieee.org/document/10023233). *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. pp. 295–301. arXiv:2202.09167 (https://arxiv.org/abs/2202.09167). doi:10.1109/SLT54892.2023.10023233 (https://doi.org/10.1109%2FSLT54892.2023.10023233). ISBN 979-8-3503-9690-4.
100. Chan, William; Jaitly, Navdeep; Le, Quoc; Vinyals, Oriol (2016). "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition" (https://storage.googleapis.com/pub-tools-public-publication-data/pdf/44926.pdf) (PDF). *ICASSP*. Archived (https://web.archive.org/web/20240909053931/https://storage.googleapis.com/pub-tools-public-publication-data/pdf/44926.pdf) (PDF) from the original on 9 September 2024. Retrieved 9 September 2024.
101. Bahdanau, Dzmitry (2016). "End-to-End Attention-based Large Vocabulary Speech Recognition". arXiv:1508.04395 (https://arxiv.org/abs/1508.04395) [cs.CL (https://arxiv.org/archive/cs.CL)].

102. Chorowski, Jan; Jaitly, Navdeep (8 December 2016). "Towards better decoding and language model integration in sequence to sequence models". arXiv:1612.02695 (https://arxiv.org/abs/1612.02695) [cs.NE (https://arxiv.org/archive/cs.NE)].
103. Chan, William; Zhang, Yu; Le, Quoc; Jaitly, Navdeep (10 October 2016). "Latent Sequence Decompositions". arXiv:1610.03035 (https://arxiv.org/abs/1610.03035) [stat.ML (https://arxiv.org/archive/stat.ML)].
104. Chung, Joon Son; Senior, Andrew; Vinyals, Oriol; Zisserman, Andrew (16 November 2016). "Lip Reading Sentences in the Wild". *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3444–3453. arXiv:1611.05358 (https://arxiv.org/abs/1611.05358). doi:10.1109/CVPR.2017.367 (https://doi.org/10.1109%2FCVPR.2017.367). ISBN 978-1-5386-0457-1. S2CID 1662180 (https://api.semanticscholar.org/CorpusID:1662180).
105. El Kheir, Yassine; et al. (21 October 2023), *Automatic Pronunciation Assessment — A Review*, Conference on Empirical Methods in Natural Language Processing, arXiv:2310.13974 (https://arxiv.org/abs/2310.13974), S2CID 264426545 (https://api.semanticscholar.org/CorpusID:264426545)
106. Isaacs, Talia; Harding, Luke (July 2017). "Pronunciation assessment" (https://doi.org/10.1017%2FS0261444817000118). *Language Teaching*. **50** (3): 347–366. doi:10.1017/S0261444817000118 (https://doi.org/10.1017%2FS0261444817000118). ISSN 0261-4448 (https://search.worldcat.org/issn/0261-4448). S2CID 209353525 (https://api.semanticscholar.org/CorpusID:209353525).
107. Loukina, Anastassia; et al. (6 September 2015), "Pronunciation accuracy and intelligibility of non-native speech" (https://www.isca-speech.org/archive/pdfs/interspeech_2015/loukina15_interspeech.pdf) (PDF), *INTERSPEECH 2015*, Dresden, Germany: International Speech Communication Association, pp. 1917–1921, archived (https://web.archive.org/web/20240909053932/https://www.isca-speech.org/archive/pdfs/interspeech_2015/loukina15_interspeech.pdf) (PDF) from the original on 9 September 2024, retrieved 9 September 2024, "only 16% of the variability in word-level intelligibility can be explained by the presence of obvious mispronunciations."
108. O'Brien, Mary Grantham; et al. (31 December 2018). "Directions for the future of technology in pronunciation research and teaching" (https://doi.org/10.1075%2Fjslp.17001.obr). *Journal of Second Language Pronunciation*. **4** (2): 182–207. doi:10.1075/jslp.17001.obr (https://doi.org/10.1075%2Fjslp.17001.obr). hdl:2066/199273 (https://hdl.handle.net/2066%2F199273). ISSN 2215-1931 (https://search.worldcat.org/issn/2215-1931). S2CID 86440885 (https://api.semanticscholar.org/CorpusID:86440885). "pronunciation researchers are primarily interested in improving L2 learners' intelligibility and comprehensibility, but they have not yet collected sufficient amounts of representative and reliable data (speech recordings with corresponding annotations and judgments) indicating which errors affect these speech dimensions and which do not. These data are essential to train ASR algorithms to assess L2 learners' intelligibility."
109. Eskenazi, Maxine (January 1999). "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype" (https://www.lltjournal.org/item/10125-25043/). *Language Learning & Technology*. **2** (2): 62–76. Archived (https://web.archive.org/web/20240909053942/https://www.lltjournal.org/item/10125-25043/) from the original on 9 September 2024. Retrieved 11 February 2023.
110. Tholfsen, Mike (9 February 2023). "Reading Coach in Immersive Reader plus new features coming to Reading Progress in Microsoft Teams" (https://techcommunity.microsoft.com/t5/education-blog/reading-coach-in-immersive-reader-plus-new-features-coming-to/ba-p/3734079). *Techcommunity Education Blog*. Microsoft. Archived (https://web.archive.org/web/20240909052822/https://techcommunity.microsoft.com/t5/education-blog/reading-coach-in-immersive-reader-plus-new-features-coming-to/ba-p/3734079) from the original on 9 September 2024. Retrieved 12 February 2023.

111. Banerji, Olina (7 March 2023). "Schools Are Using Voice Technology to Teach Reading. Is It Helping?" (https://www.edsurge.com/news/2023-03-07-schools-are-using-voice-technology-to-teach-reading-is-it-helping). *EdSurge News*. Archived (https://web.archive.org/web/20240909054611/https://www.edsurge.com/news/2023-03-07-schools-are-using-voice-technology-to-teach-reading-is-it-helping) from the original on 9 September 2024. Retrieved 7 March 2023.
112. Hair, Adam; et al. (19 June 2018). "Apraxia world: A speech therapy game for children with speech sound disorders". *Proceedings of the 17th ACM Conference on Interaction Design and Children* (https://psi.engr.tamu.edu/wp-content/uploads/2018/04/hair2018idc.pdf) (PDF). pp. 119–131. doi:10.1145/3202185.3202733 (https://doi.org/10.1145%2F3202185.3202733). ISBN 9781450351522. S2CID 13790002 (https://api.semanticscholar.org/CorpusID:13790002). Archived (https://web.archive.org/web/20240909052803/https://psi.engr.tamu.edu/wp-content/uploads/2018/04/hair2018idc.pdf) (PDF) from the original on 9 September 2024. Retrieved 9 September 2024.
113. "Computer says no: Irish vet fails oral English test needed to stay in Australia" (https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia). *The Guardian*. Australian Associated Press. 8 August 2017. Archived (https://web.archive.org/web/20240909052806/https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia) from the original on 9 September 2024. Retrieved 12 February 2023.
114. Ferrier, Tracey (9 August 2017). "Australian ex-news reader with English degree fails robot's English test" (https://www.smh.com.au/technology/australian-exnews-reader-with-english-degree-fails-robots-english-test-20170809-gxsjv2.html). *The Sydney Morning Herald*. Archived (https://web.archive.org/web/20240909053307/https://www.smh.com.au/technology/australian-exnews-reader-with-english-degree-fails-robots-english-test-20170809-gxsjv2.html) from the original on 9 September 2024. Retrieved 12 February 2023.
115. Main, Ed; Watson, Richard (9 February 2022). "The English test that ruined thousands of lives" (https://www.bbc.com/news/uk-60264106). *BBC News*. Archived (https://web.archive.org/web/20240909054614/https://www.bbc.com/news/uk-60264106) from the original on 9 September 2024. Retrieved 12 February 2023.
116. Joyce, Katy Spratte (24 January 2023). "13 Words That Can Be Pronounced Two Ways" (https://www.rd.com/list/words-that-can-be-pronounced-two-ways/). Reader's Digest. Archived (https://web.archive.org/web/20240909054447/https://www.rd.com/list/words-that-can-be-pronounced-two-ways/) from the original on 9 September 2024. Retrieved 23 February 2023.
117. E.g., CMUDICT, "The CMU Pronouncing Dictionary" (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). *www.speech.cs.cmu.edu*. Archived (https://web.archive.org/web/20100815023012/http://www.speech.cs.cmu.edu/cgi-bin/cmudict) from the original on 15 August 2010. Retrieved 15 February 2023. Compare "four" given as "F AO R" with the vowel AO as in "caught," to "row" given as "R OW" with the vowel OW as in "oat."
118. Tu, Zehai; Ma, Ning; Barker, Jon (2022). "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction" (https://www.isca-speech.org/archive/pdfs/interspeech_2022/tu22b_interspeech.pdf) (PDF). *Proc. Interspeech 2022*. INTERSPEECH 2022. ISCA. pp. 3493–3497. doi:10.21437/Interspeech.2022-10408 (https://doi.org/10.21437%2FInterspeech.2022-10408). Archived (https://web.archive.org/web/20240909053824/https://www.isca-speech.org/archive/pdfs/interspeech_2022/tu22b_interspeech.pdf) (PDF) from the original on 9 September 2024. Retrieved 17 December 2023.
119. *Common European framework of reference for languages learning, teaching, assessment: Companion volume with new descriptors* (https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989). Language Policy Programme, Education Policy Division, Education Department, Council of Europe. February 2018. p. 136. OCLC 1090351600 (https://search.worldcat.org/oclc/1090351600). Archived (https://web.archive.org/web/20240909053825/https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989) from the original on 9 September 2024. Retrieved 9 September 2024.

120. Englund, Christine (2004). *Speech recognition in the JAS 39 Gripen aircraft: Adaptation to speech at different G-loads* (http://www.speech.kth.se/prod/publications/files/1664.pdf) (PDF) (Masters thesis thesis). Stockholm Royal Institute of Technology. Archived (https://web.archive.org/web/20081002002102/http://www.speech.kth.se/prod/publications/files/1664.pdf) (PDF) from the original on 2 October 2008.
121. "The Cockpit" (https://www.eurofighter.com/the-aircraft#cockpit). *Eurofighter Typhoon*. Archived (https://web.archive.org/web/20170301222529/https://www.eurofighter.com/the-aircraft#cockpit) from the original on 1 March 2017.
122. "Eurofighter Typhoon – The world's most advanced fighter aircraft" (http://www.eurofighter.com/capabilities/technology/voice-throttle-stick/direct-voice-input.html). *www.eurofighter.com*. Archived (https://web.archive.org/web/20130511025203/http://www.eurofighter.com/capabilities/technology/voice-throttle-stick/direct-voice-input.html) from the original on 11 May 2013. Retrieved 1 May 2018.
123. Schutte, John (15 October 2007). "Researchers fine-tune F-35 pilot-aircraft speech system" (https://web.archive.org/web/20071020030310/http://www.af.mil/news/story.asp?id=123071861). United States Air Force. Archived from the original (http://www.af.mil/news/story.asp?id=123071861) on 20 October 2007.
124. "Overcoming Communication Barriers in the Classroom" (http://www.massmatch.org/aboutus/listserv/2010/2010-03-31.html). MassMATCH. 18 March 2010. Archived (https://web.archive.org/web/20130725024622/http://www.massmatch.org/aboutus/listserv/2010/2010-03-31.html) from the original on 25 July 2013. Retrieved 15 June 2013.
125. "Speech Recognition for Learning" (http://www.brainline.org/content/2010/12/speech-recognition-for-learning_pageall.html). National Center for Technology Innovation. 2010. Archived (https://web.archive.org/web/20140413100513/http://www.brainline.org/content/2010/12/speech-recognition-for-learning_pageall.html) from the original on 13 April 2014. Retrieved 26 March 2014.
126. Follensbee, Bob; McCloskey-Dale, Susan (2000). "Speech recognition in schools: An update from the field" (http://www.csun.edu/~hfdss006/conf/2000/proceedings/0219Follansbee.htm). *Technology And Persons With Disabilities Conference 2000*. Archived (https://web.archive.org/web/20060821213145/http://www.csun.edu/~hfdss006/conf/2000/proceedings/0219Follansbee.htm) from the original on 21 August 2006. Retrieved 26 March 2014.
127. "Speech recognition for disabled people" (https://web.archive.org/web/20080404013302/http://www.businessweek.com/1998/08/b3566022.htm). Archived from the original (http://www.businessweek.com/1998/08/b3566022.htm) on 4 April 2008.
128. Friends International Support Group
129. Garrett, Jennifer Tumlin; et al. (2011). "Using Speech Recognition Software to Increase Writing Fluency for Individuals with Physical Disabilities" (https://scholarworks.gsu.edu/epse_diss/46). *Journal of Special Education Technology*. **26** (1): 25–41. doi:10.1177/016264341102600104 (https://doi.org/10.1177%2F016264341102600104). S2CID 142730664 (https://api.semanticscholar.org/CorpusID:142730664). Archived (https://web.archive.org/web/20240909053848/https://scholarworks.gsu.edu/epse_diss/46/) from the original on 9 September 2024. Retrieved 9 September 2024.
130. Forgrave, Karen E. "Assistive Technology: Empowering Students with Disabilities." Clearing House 75.3 (2002): 122–6. Web.
131. Tang, K. W.; Kamoua, Ridha; Sutan, Victor (2004). "Speech Recognition Technology for Disabilities Education". *Journal of Educational Technology Systems*. **33** (2): 173–84. CiteSeerX 10.1.1.631.3736 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.631.3736). doi:10.2190/K6K8-78K2-59Y7-R9R2 (https://doi.org/10.2190%2FK6K8-78K2-59Y7-R9R2). S2CID 143159997 (https://api.semanticscholar.org/CorpusID:143159997).
132. "Projects: Planetary Microphones" (https://web.archive.org/web/20120127161038/http://www.planetary.org/programs/projects/planetary_microphones/mars_microphone.html). The Planetary Society. Archived from the original (http://www.planetary.org/programs/projects/planetary_microphones/mars_microphone.html) on 27 January 2012.

133. Caridakis, George; Castellano, Ginevra; Kessous, Loic; Raouzaiou, Amaryllis; Malatesta, Lori; Asteriadis, Stelios; Karpouzis, Kostas (19 September 2007). "Multimodal emotion recognition from expressive faces, body gestures and speech". *Artificial Intelligence and Innovations 2007: From Theory to Applications*. IFIP the International Federation for Information Processing. Vol. 247. Springer US. pp. 375–388. doi:10.1007/978-0-387-74161-1_41 (https://doi.org/10.1007%2F978-0-387-74161-1_41). ISBN 978-0-387-74160-4.
134. "What is real-time captioning? | DO-IT" (https://www.washington.edu/doit/what-real-time-captioning). *www.washington.edu*. Archived (https://web.archive.org/web/20240909054510/https://www.washington.edu/doit/what-real-time-captioning) from the original on 9 September 2024. Retrieved 11 April 2021.
135. Zheng, Thomas Fang; Li, Lantian (2017). *Robustness-Related Issues in Speaker Recognition* (http://link.springer.com/10.1007/978-981-10-3238-7). SpringerBriefs in Electrical and Computer Engineering. Singapore: Springer Singapore. doi:10.1007/978-981-10-3238-7 (https://doi.org/10.1007%2F978-981-10-3238-7). ISBN 978-981-10-3237-0. Archived (https://web.archive.org/web/20240909053948/https://link.springer.com/book/10.1007/978-981-10-3238-7) from the original on 9 September 2024. Retrieved 9 September 2024.
136. Ciaramella, Alberto. "A prototype performance evaluation report." Sundial workpackage 8000 (1993).
137. Gerbino, E.; Baggia, P.; Ciaramella, A.; Rullent, C. (1993). "Test and evaluation of a spoken dialogue system". *IEEE International Conference on Acoustics Speech and Signal Processing*. pp. 135–138 vol.2. doi:10.1109/ICASSP.1993.319250 (https://doi.org/10.1109%2FICASSP.1993.319250). ISBN 0-7803-0946-4. S2CID 57374050 (https://api.semanticscholar.org/CorpusID:57374050).
138. National Institute of Standards and Technology. "The History of Automatic Speech Recognition Evaluation at NIST (http://www.itl.nist.gov/iad/mig/publications/ASRhistory/) Archived (https://web.archive.org/web/20131008210040/http://www.itl.nist.gov/iad/mig/publications/ASRhistory/) 8 October 2013 at the Wayback Machine".
139. "Letter Names Can Cause Confusion and Other Things to Know About Letter–Sound Relationships" (https://www.naeyc.org/resources/pubs/yc/mar2015/letter-sound-relationships). *NAEYC*. Archived (https://web.archive.org/web/20240909054452/https://www.naeyc.org/resources/pubs/yc/mar2015/letter-sound-relationships) from the original on 9 September 2024. Retrieved 27 October 2023.
140. "Listen Up: Your AI Assistant Goes Crazy For NPR Too" (https://www.npr.org/2016/03/06/469383361/listen-up-your-ai-assistant-goes-crazy-for-npr-too). *NPR*. 6 March 2016. Archived (https://web.archive.org/web/20170723210358/http://www.npr.org/2016/03/06/469383361/listen-up-your-ai-assistant-goes-crazy-for-npr-too) from the original on 23 July 2017.
141. Claburn, Thomas (25 August 2017). "Is it possible to control Amazon Alexa, Google Now using inaudible commands? Absolutely" (https://www.theregister.co.uk/2017/08/25/amazon_alexa_answers_inaudible_commands/?mt=1504024969000). *The Register*. Archived (https://web.archive.org/web/20170902051123/https://www.theregister.co.uk/2017/08/25/amazon_alexa_answers_inaudible_commands/?mt=1504024969000) from the original on 2 September 2017.
142. "Attack Targets Automatic Speech Recognition Systems" (https://motherboard.vice.com/en_us/article/d34nnz/attack-targets-automatic-speech-recognition-systems). *vice.com*. 31 January 2018. Archived (https://web.archive.org/web/20180303050744/https://motherboard.vice.com/en_us/article/d34nnz/attack-targets-automatic-speech-recognition-systems) from the original on 3 March 2018. Retrieved 1 May 2018.
143. Beigi, Homayoon (2011). *Fundamentals of Speaker Recognition* (http://www.fundamentalsofspeakerrecognition.org). New York: Springer. ISBN 978-0-387-77591-3. Archived (https://web.archive.org/web/20180131140911/http://www.fundamentalsofspeakerrecognition.org/) from the original on 31 January 2018.
144. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society.

145. "Common Voice by Mozilla" (https://web.archive.org/web/20200227020208/https://voice.mozilla.org/). *voice.mozilla.org*. Archived from the original (https://voice.mozilla.org/) on 27 February 2020. Retrieved 9 November 2019.
146. "A TensorFlow implementation of Baidu's DeepSpeech architecture: mozilla/DeepSpeech" (https://github.com/mozilla/DeepSpeech). 9 November 2019. Archived (https://web.archive.org/web/20240909053949/https://github.com/mozilla/DeepSpeech) from the original on 9 September 2024. Retrieved 9 September 2024 – via GitHub.
147. "GitHub - tensorflow/docs: TensorFlow documentation" (https://github.com/tensorflow/docs). 9 November 2019. Archived (https://web.archive.org/web/20240909053830/https://github.com/tensorflow/docs) from the original on 9 September 2024. Retrieved 9 September 2024 – via GitHub.
148. "Coqui, a startup providing open speech tech for everyone" (https://github.com/coqui-ai). *GitHub*. Archived (https://web.archive.org/web/20240909054614/https://github.com/coqui-ai) from the original on 9 September 2024. Retrieved 7 March 2022.
149. Coffey, Donavyn (28 April 2021). "Māori are trying to save their language from Big Tech" (https://www.wired.co.uk/article/maori-language-tech). *Wired UK*. ISSN 1357-0978 (https://search.worldcat.org/issn/1357-0978). Archived (https://web.archive.org/web/20240909053950/https://www.wired.com/story/maori-language-tech/) from the original on 9 September 2024. Retrieved 16 October 2021.
150. "Why you should move from DeepSpeech to coqui.ai" (https://discourse.mozilla.org/t/why-you-should-move-from-deepspeech-to-coqui-ai/82798). *Mozilla Discourse*. 7 July 2021. Retrieved 16 October 2021.
151. "Type with your voice" (https://support.google.com/gboard/answer/2781851?hl=en&co=GENIE.Platform%3DAndroid). Archived (https://web.archive.org/web/20240909054332/https://support.google.com/gboard/answer/2781851?hl=en&co=GENIE.Platform%3DAndroid) from the original on 9 September 2024. Retrieved 9 September 2024.

Further reading

- Cole, Ronald; Mariani, Joseph; Uszkoreit, Hans; Varile, Giovanni Battista; Zaenen, Annie; Zampolli; Zue, Victor, eds. (1997). *Survey of the state of the art in human language technology*. Cambridge Studies in Natural Language Processing. Vol. XII–XIII. Cambridge University Press. ISBN 978-0-521-59277-2.
- Junqua, J.-C.; Haton, J.-P. (1995). *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers. ISBN 978-0-7923-9646-8.
- Karat, Clare-Marie; Vergo, John; Nahamoo, David (2007). "Conversational Interface Technologies". In Sears, Andrew; Jacko, Julie A. (eds.). *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications (Human Factors and Ergonomics)*. Lawrence Erlbaum Associates Inc. ISBN 978-0-8058-5870-9.
- Pieraccini, Roberto (2012). *The Voice in the Machine. Building Computers That Understand Speech*. The MIT Press. ISBN 978-0262016858.
- Pirani, Giancarlo, ed. (2013). *Advanced algorithms and architectures for speech understanding*. Springer Science & Business Media. ISBN 978-3-642-84341-9.
- Signer, Beat; Hoste, Lode (December 2013). "SpeeG2: A Speech- and Gesture-based Interface for Efficient Controller-free Text Entry" (https://www.academia.edu/4685517). *Proceedings of ICMI 2013*. 15th International Conference on Multimodal Interaction. Sydney, Australia.

- Woelfel, Matthias; McDonough, John (26 May 2009). *Distant Speech Recognition*. Wiley. ISBN 978-0470517048.

External links

- Speech Technology (https://curlie.org/Computers/Speech_Technology) at Curlie

Retrieved from "https://en.wikipedia.org/w/index.php?title=Speech_recognition&oldid=1248962061"