REVIEW ARTICLES

A Historical Perspective of Speech Recognition

By Xuedong Huang, James Baker, Raj Reddy

Comments

PRINT

VIEW AS: SHARE: G+

Credit: Shutterstock.com

With the introduction of Apple's Siri and similar voice search services from Google and Microsoft, it is natural to wonder why it has taken so long for voice recognition technology to advance to this level. Also, we wonder, when can we expect to hear a more human-level performance? In 1976, one of the authors (Reddy) wrote a comprehensive review of the state of the art of voice recognition at that time. A non-expert in the field may benefit from reading the original article.[34] Here, we provide our collective historical perspective on the advances in the field of speech recognition. Given the space limitations, this article will not attempt a comprehensive technical review, but limit the scope to discussing the missing science of speech recognition 40 years ago and what advances seem to have contributed to overcoming some of the most thorny problems.

Back to Top

Key Insights

A Historical Perspective of Speech Recognition

de CACM

03:43

A Historical Perspective of Speech Recognition from CACM on Vimeo.

- The insights gained from the speech recognition advances over the past 40 years are explored, originating from generations of Carnegie Mellon University's R&D.
- Several major achievements over the years have proven to work well in practice for leading industry speech recognition systems from Apple to Microsoft.
- Speech recognition will pass the Turing Test and bring the vision of Star Trek-like mobile devices to reality. It will help to bridge the gap between humans and machines. It will facilitate and enhance natural conservation among people. Six challenges need to be addressed before we can realize this audacious dream.

Speech recognition had been a staple of science fiction for years, but in 1976 the real-world capabilities bore little resemblance to the far-fetched capabilities in the fictional realm. Nonetheless, Reddy boldly predicted it would be possible to build a $20,000 connected speech system within the next 10 years. Although it took longer than projected,

MORE NEWS & OPINIONS

Mathematics Can Assist Cities in Addressing Unstructured Neighborhoods
Oak Ridge National Laboratory

What the Boston School Bus Schedule Can Teach Us About AI
Wired

Tech User Responsibility
Robin K. Hill

ACM RESOURCES

Wireless LAN Solutions v1.0
Courses

A Historical Perspective of Speech Recognition from CACM on Vimeo.

">

not only were the goals eventually met, but the system costs were much less and have continued to drop dramatically. Today, in many smartphones, the industry delivers free speech recognition that significantly exceeds Reddy's speculations. In most fields the imagination of science fiction writers far exceeds reality. Speech recognition is one of the few exceptions. Moreover, speech recognition is unique not just because of its successes: in spite of all the accomplishments, additional challenges remain that are as daunting as those that have been overcome to date.

In 1995, Microsoft SAPI was first shipped in Windows 95 to enable application developers to create speech applications on Windows. In 1999 the VoiceXML forum was created to support telephony IVR. While speech-enabled telephony IVR was commercially successful, it has been shown the "speech in" and "screen out" multimodal metaphor is more natural for information consumption. In 2001, Bill Gates demonstrated such a prototype codenamed MiPad at CES.[16] MiPad illustrated a vision on speech-enabled multimodal mobile devices. With the recent adoption of speech recognition used in Apple, Google, and Microsoft products, we are witnessing the ever-improved ability of devices to handle relatively unrestricted multimodal dialogues. We see the fruits of several decades of R&D in spite of remaining challenges. We believe the speech community is en route to pass the Turing Test in the next 40 years with the ultimate goal to match and exceed a human's speech recognition capability for everyday scenarios.

Here, we highlight major speech recognition technologies that worked well in practice and summarize six challenging areas that are critical to move speech recognition to the next level from the current showcase services on mobile devices. More comprehensive technical discussions may be found in the numerous technical papers published over the last decade, including *IEEE Transactions on Audio, Speech and Language Processing* and *Computer Speech and Language*, as well as proceedings from ICASSP, Interspeech, and IEEE workshops on ASRU. There are also numerous articles and books that cover systems and technologies developed over the last four decades.[9,14,15,19,25,33,36,43]

Back to Top

Basic Speech Recognition

In 1971, a speech recognition study group chaired by Allen Newell recommended that many more sources of knowledge be brought to bear on the problem. The report discussed six levels of knowledge: acoustic, parametric, phonemic, lexical, sentence, and semantic. Klatt[23] provides a review of performance of various ARPA-funded speech understanding systems initiated to achieve the goals of Newell report.

By 1976, Reddy was leading a group at Carnegie Mellon University that was one of a small number of research groups funded to explore the ideas in the Newell report under a multiyear Defense Advanced Research Project Agency (DARPA)-sponsored Speech Understanding Research (SUR) project. This group developed a sequence of speech recognition systems: Hearsay, Dragon, Harpy, and Sphinx I/II. Over a span of four decades, Reddy and his colleagues created several historic demonstrations of spoken language systems, for example, voice control of a robot, large-vocabulary connected-speech recognition, speaker-independent speech recognition, and unrestricted vocabulary dictation. Hearsay-I was one of the first systems capable of continuous speech recognition. The Dragon system was one of the first systems to model speech as a hidden stochastic process. The Harpy system introduced the concept of Beam Search, which for decades has been the most widely used technique for efficient searching and matching. Sphinx-I, developed in 1987, was the first system to demonstrate speaker-independent speech recognition. Sphinx-II, developed in 1992, benefited largely from tied parameters to balance trainability and efficiency at both Gaussian mixture and Markov state level, which achieved the highest recognition accuracy in DARPA-funded speech benchmark evaluation in 1992.

As per the DARPA-funded speech evaluations, the speech recognition word error rate has been used as the main metric to evaluate the progress. The historical progress also directed the community to work on more difficult speech recognition tasks as shown in Figure 1. On the latest switchboard task, the word error rate is approaching an impressive new milestone by both Microsoft and IBM researchers respectively,[4,22,37] following the deep learning framework pioneered by researchers at the University of Toronto and Microsoft.[5,14]

It was anticipated in the early 1970s that to bring to bear the higher-level sources of knowledge might require significant breakthroughs in artificial intelligence. The architecture of the Hearsay system was designed so that many semiautonomous modules can communicate and cooperate in a speech recognition task while each concentrated on its own area of expertise. In contrast, the Dragon, Harpy, and Sphinx I/II systems were all based on a single, relatively simple modeling principle of joint global optimization. Each of the levels in the Newell report was represented by a stochastic process known as a hidden Markov process. Successive levels were conceptually embedded like nesting blocks, so the combined process was also a (very large) hidden Markov process.[2]

The decoding process of finding the best matched word sequence W to match input speech X is more than a simple pattern recognition problem, since one faces a practically astronomical number of word patterns to search. The decoding process in a speech recognizer's operation is to find a sequence of words whose corresponding acoustic and language models best match the input feature vector sequence. Thus, such a decoding process with trained acoustic and language models is often referred to as a search process. Graph search algorithms, which have been explored extensively in the fields of artificial intelligence, operations research, and game theory, serve as the basic foundation for the search problem in speech recognition.

The importance of the decoding process is best illustrated by Dragon NaturallySpeaking, a product that took 15 years to develop under the leadership of one of the authors (Baker). It has survived for 15 years through many generations of computer technology after being acquired by Nuance. Dragon Systems did not owe its success to inventing radically new algorithms with superior performance. The development of technology for Dragon NaturallySpeaking may be compared with the general development in the same timeframe reviewed in this article. The most salient difference is not algorithms with a lower error rate, but rather an emphasis on simplified algorithms with a better cost-performance trade-off. From its founding, the long-term goal of Dragon Systems was the development of a real-time, large-vocabulary, continuous-speech dictation system. Toward that end, Dragon formulated a coherent mission statement that would last

for decades that would be required to reach the long-term goal, but that in each time frame would translate into appropriate short-term and medium-term objectives: Produce the best speech recognition that could run in real time on the current generation of desktop computers.

What We Did Not Know in 1976

Each of the components illustrated in Reddy's original review paper has made significant progress. We do not plan to enumerate all the different systems and approaches developed over the decades. Table 1 contains the major achievements that are proven to work well in practice for leading industry speech recognition systems. Today, we can use open research tools, such as HTK, Sphinx, Kaldi, CMU LM toolkit, and SRILM to build a working system. However, the competitive edge in the industry mostly benefited from using a massive amount of data available in the cloud to continuously update and improve the acoustic model and the language model. Here, we discuss progress that enabled today's voice search on mobile phones such as Apple, Google, and Microsoft Voice Search as illustrated in Figure 2.

The establishment of the statistical machine-learning framework, supported by the availability of computing infrastructure and massive training data, constitutes the most significant driving force in advancing the development of speech recognition. This enabled machine learning to treat phonetic, word, syntactic, and semantic knowledge representations in a unified manner. For example, explicit segmentation and labeling of phonetic strings is no longer necessary. Phonetic matching and word verification are unified with word sequence generation that depends on the highest overall rating typically using a context-dependent phonetic acoustic model.

Statistical machine learning. Early methods of speech recognition aimed to find the closest matching sound label from a discrete set of labels. In non-probabilistic models, there is an estimated "distance" between sound labels based on how similar two sounds are estimated to be. In one form, probability models use an estimate of the conditional probability of observing a particular sound label as the best matching label, conditional on the correct label being the hypothesized label, which is also called the "confusion" probability. To estimate the probability of confusing each possible sound with each possible label requires substantially more training data than estimating the mean of a Gaussian distribution, another common representation. This method corresponds to the "labeling" part of the "segmentation and labeling" described in Reddy's 1976 review, whether accompanied by segmentation or not, as was often done by the 1980s for non-probability-based models. This distance may merely be a score to be minimized.

A pivotal change in the representation of knowledge in speech recognition was just beginning at the time of Reddy's review paper. This change was exemplified by the representation of speech as a hidden Markov process. This is usually referred to with the acronym HMM for "Hidden Markov Model," which is a slight misnomer because it is the process that is hidden not the model.[2] Mathematically, the model for a hidden Markov process has a learning algorithm with a broadly applicable convergence theorem called the Expectation-Maximization (EM) algorithm.[3,8] In the particular case of a hidden Markov process, it has a very efficient implementation via the Forward-Backward algorithm. Since the late 1980s, statistical discriminative training techniques have also been developed based on maximum mutual information or related minimum error criteria.[1,13,21]

Before 2010, a mixture of HMM-based Gaussian densities have typically been used for state-of-the-art speech recognition. The features for these models are typically Mel-frequency cepstral coefficients (MFCC).[6] While there are many efforts in creating features imitating the human auditory process, we want to highlight one significant development that offers learned feature representation with the introduction of deep neural networks (DNN). Overcoming the inefficiency in data representation by the Gaussian mixture model, DNN can replace the Gaussian mixture model directly.[14] Deep learning can also be used to learn powerful discriminative features for a traditional HMM speech recognition system.[37] The advantage of this hybrid system is that decades of speech recognition technologies developed by speech recognition researchers can be used directly. A combination of DNN and HMM produced significant error reduction[4,14,22,37] in comparison to some of the early efforts.[29,40] In the new system, the speech classes for DNN are typically represented by tied HMM states—a technique directly inherited from earlier speech systems.[18]

Using Markov models to represent language knowledge was controversial. Linguists knew no natural language could be represented even by context-free grammar, much less by a finite state grammar. Similarly, artificial intelligence experts were more doubtful that a model as simple as a Markov process would be useful for representing the higher-level knowledge sources recommended in the Newell report.

However, there is a fundamental difference between assuming that language itself is a Markov process and modeling language as a probabilistic function of a hidden Markov process. The latter model is an approximation method that does not make an assumption about language, but rather provides a prescription to the designer in choosing what to represent in the hidden process. The definitive property of a Markov process is that, given the current state, probabilities of future events will be independent of any additional information about the past history of the process. This property means if there is any information about the past history of the observed process (such as the observed words and sub-word units), then the designer should encode that information with distinct states in the hidden process. It turned out that each of the levels of the Newell hierarchy could be represented as a probabilistic function of a hidden Markov process to a reasonable level of approximation.

For today's state-of-the-art language modeling, most systems still use the statistical *N*-gram language models and the variants, trained with the basic counting or EM-style techniques. These models have proved remarkably powerful and resilient. However, the *N*-gram is a highly simplistic model for realistic human language. In a similar manner with deep learning for significantly improving acoustic modeling quality, recurrent neural networks have also significantly improved the *N*-gram language model.[27] It is worth noting that nothing beats a massive text corpora matching the application domain for most real speech applications.

Training data and computational resources. The availability of speech/text data and computing power has been instrumental in enabling speech recognition researchers to develop and evaluate complex algorithms on sufficiently large tasks. The availability of common speech corpora for speech training, development, and evaluation, has been critical, allowing the creation of complex systems of ever-increasing capabilities. Since speech is a highly variable signal and is characterized by many parameters, large corpora become critical in modeling it well enough for automated systems to achieve proficiency. Over the years, these corpora have been created, annotated, and distributed to the worldwide community by the National Institute of Standard and Technology (NIST), the Linguistic Data Consortium (LDC), European Language Resources Association (ELRA), and other organizations. The character of the recorded speech has progressed from limited, constrained speech materials to huge amounts of progressively more realistic, spontaneous speech.

Moore's Law predicts doubling the amount of computation for a given cost every 12–18 months, as well as a comparably shrinking cost of memory. Moore's Law made it possible for speech recognition to consume the significantly improved computational infrastructure. Cloud-based speech recognition made it more convenient to accumulate an even more massive amount of speech data than ever imagined in 1976. Both Google and Bing indexed the entire Web. Billions of user queries reach the Web search engine monthly. This massive amount of query click data made it possible to create a far more powerful language model for voice search applications.

Signal and feature processing. A vector of acoustic features is computed typically every 10 milliseconds. For each frame a short window of speech data is selected. Typically each window selects about 25 milliseconds of speech, so the windows overlap in time. In 1976, the acoustic features were typically a measure of the magnitude at each of a set of frequencies for each time window, typically computed by a fast Fourier transform or by a filter bank. The magnitude as function of frequency is called the "spectrum" of the short time window of speech, and a sequence of such spectra over time in a speech utterance can be visualized as a spectrogram.[31]

Over the past 30 years or so, modifications of spectrograms led to significant improvements in the performance of Gaussian mixture-based HMM systems despite the loss of raw speech information due to such modifications. Deep learning technology aims squarely at minimizing such information loss and at searching for more powerful, deep learning-driven speech representations from raw data. As a result of the success in deep learning, speech recognition researchers are returning to using more basic speech features such as spectrograms and filterbanks for deep learning,[11] allowing the power of machine learning to automatically discover more useful representations from the DNN itself.[37,39]

Vocabulary size. The maximum vocabulary size for large speech recognition has increased substantially since 1976. In fact, for real-time natural language dictation systems in the late 1990s the vocabulary size essentially became unlimited. That is, the user was not aware of which relatively rare words were in the system's dictionary and which were not. The systems tried to recognize every word dictated and counted as an error any word that was not recognized, even if the word was not in the dictionary.

This point of view forced these systems to learn new words on the fly so the system would not keep making the same mistake every time the same word occurred. It was especially important to learn the names of people and places that occurred repeatedly in a particular user's dictation. Significant advances were made in statistical learning techniques for learning from a single example or a small number of examples. The process was made to appear as seamless as possible to the interactive user. However, the problem remains a challenge because modeling new words is still far from seamless when seen from the point of view of the models, where the small-sample models are quite different from the large-data models.

Speaker independent and adaptive systems. Although probability models with statistical machine learning provided a means to model and learn many sources of variability in the speech signal, there was still a significant gap in performance between single-speaker, speaker-dependent models and speaker-independent models intended for the diverse population. Sphinx introduced large vocabulary speaker-independent continuous speech recognition.[24] The key was to use more speech data from a large number of speakers to train the HMM-based system.

Speech recognition is unique not just because of its successes: in spite of all the accomplishments, additional challenges remain that are as daunting as those that have been overcome so far.

Adaptive learning is also applied to accommodate speaker variations and a wide range of variable conditions for the channel, noise, and domain.[24] Effective adaptation technologies enable rapid application integration, and are a key to successful commercial deployment of speech recognition.

Decoding techniques. Architecturally, the most important development in knowledge representation has been searchable unified graph representations that allow multiple sources of knowledge to be incorporated into a common probabilistic framework. The decoding or search strategies have evolved from many systems summarized in Reddy's 1976 paper, such as stack decoding (A* search),[20] time-synchronous beam search,[26] and Weighted Finite State Transducer (WFST) decoder.[28] These practical decoding algorithms made possible large-scale continuous speech recognition.

Non-compositional methods include multiple speech streams, multiple probability estimators, multiple recognition systems combined at the hypothesis level such as ROVER,[12] and multi-pass systems with increased constraints.

Spoken language understanding. Once recognition results are available, it is equally important to extract "meaning" for the recognition results. Spoken language understanding (SLU) mostly relied on case grammars for representing sets of

semantic concepts during 1970s. A good example of putting the case grammars for SLU is exemplified by the Air Travel Information System (ATIS) research initiative funded by DARPA.[32,41] In this task, the users can utter queries on flight information in an unrestricted free form. Understanding the spoken language is about extracting task-specific arguments in a given frame-based semantic representation involving frames such as "departure date," and "flight." The slot in these case frames is specific to the domain involved. Finding the value of properties from speech recognition results must be robust to deal with inherent recognition errors as well as a wide range of different ways of expressing the same concept.

A number of techniques are used to fill frame slots of the application domain from the training data.[30,35,41] Like acoustic and language modeling, deep learning based on recurrent neural networks can also significantly improve filling slots for language understanding.[38]

Six Major Challenges

Speech recognition technology is far from perfect. Indeed, technical challenges abound. Based on what we have learned over the past 40 years, we now discuss six of the most challenging areas to be addressed before we can realize the dream of speech recognition.

There is no data like more data. Today we have some very exciting opportunities to collect large amounts of data, thus giving rise to "data deluge." Thanks in large part to the Internet, there are now readily accessible large quantities of everyday speech, reflecting a variety of materials and environments previously unavailable. Recently emerging voice search in mobile phones has provided a rich source of speech data, which, because of the recording of mobile phone users' actions, can be considered as partially "labeled." Apple Siri (powered by Nuance), Google, and Microsoft all have accumulated a massive amount of user data in using voice systems on their products.

New Web-based tools could be made available to collect, annotate, and process substantial quantities of speech in a cost-effective manner in many languages. Mustering the assistance of interested individuals on the Web could generate substantial quantities of language resources very efficiently and cost effectively. This could be especially valuable for creating significant new capabilities for resource "impoverished" languages.

The ever-increasing amount of data presents both an opportunity and a challenge for advancing the state of the art in speech recognition as illustrated in Figure 3, in which our Microsoft colleagues Li Deng and Eric Horvitz used the data from a number of published papers to illustrate the key point. The numbers in Figure 3 are not precise even with our best effort to derive a cohesive chart from data scattered over a period of approximately 10 years.

We have barely scratched the surface in sampling the many kinds of speech, environments, and channels that people routinely experience. In fact, we currently provide to our automatic systems only a very small fraction of the amount of materials that humans utilize to acquire language. If we want our systems to be more powerful and to understand the nature of speech itself, we need to make more use of it and label more of it. Well-labeled speech corpora have been the cornerstone on which today's systems have been developed and evolved. However, most of the large quantities of data are not labeled or poorly "labeled," and labeling them accurately is costly.

Computing infrastructure. The use of GPUs[5,14] is a significant advancement in recent years that makes the training of modestly sized deep networks practical. A known limitation of the GPU approach is the training speed-up is small when the model does not fit in GPU memory (typically less than six gigabytes). It is recently reported that distributed optimization approach can greatly accelerate deep learning as well as enabling training larger models.[7] A cluster of massive distributed machines has been used to train a modestly sized speech DNN leading to over 10x acceleration in comparison to the GPU implementation.

Moore's Law has been a dependable indicator of the increased capability for computation and storage in our computational systems for decades. The resulting effects on systems for speech recognition and understanding have been enormous, permitting the use of larger and larger training databases and recognition systems, and the incorporation of more detailed models of spoken language. Many of the future research directions and applications implicitly depend upon continued advances in computational capabilities, which seems justified given the recent progress of using distributed computer systems to train large-scale DNNs. With the ever-increased amount of training data as illustrated in Figure 3, it is expected to take weeks or months to train a modern speech system even with a massively distributed computing cluster.

As Intel and others have recently noted, the power density on microprocessors has increased to the point that higher clock rates would begin to melt the silicon. Consequently, industry development is currently focused on implementing microprocessors on multiple cores. The new road maps for the semiconductor industry reflect this trend, and future speed-ups will come more from parallelism than from having faster individual computing elements.

For the most part, algorithm designers for speech systems have ignored investigation of such parallelism, partly because the advancement of scalability has been so reliable. Future research directions and applications will require significantly more computation resources for creating models, and consequently researchers will need to consider massive distributed parallelism in their designs. This will be a significant change from the status quo. In particular, tasks such as decoding, for which extremely clever schemes to speed up single-processor performance have been developed, will require a complete rethinking of the algorithms. New search methods that explicitly exploit parallelism should be an important research direction.

Unsupervised learning has been successfully used to train a deep network 30-times larger than previously reported.[7] With supervised fine-tuning to get the labels, DNN-based system achieved state-of-the-art performance on ImageNet, a very difficult visual object recognition task. For speech recognition, there is also a practical need to develop high-quality

unsupervised or semi-supervised techniques with a massive amount of user interaction data available in the cloud such as click data in the Web search engine.

Upon the successful development of voice search, exploitation of unlabeled or partially labeled data becomes feasible to train the underlying acoustic and language models. We can automatically (and "actively") select parts of the unlabeled data for manual labeling in a way that maximizes its utility. An important reason for unsupervised learning is the systems, like their human "baseline," will have to undergo "lifelong learning," adjusting to evolving vocabulary, channels, language use, among others. There is a need for learning at all levels to cope with changing environments, speakers, pronunciations, dialects, accents, words, meanings, and topics. Like its human counterpart, the system would engage in automatic pattern discovery, active learning, and adaptation.

We must address both the learning of new models and the integration of such models into existing systems. Thus, an important aspect of learning is being able to discern when something has been learned and how to apply the result. Learning from multiple concurrent modalities may also be necessary. For instance, a speech recognition system may encounter a new proper noun in its input speech, and may need to examine textual contexts to determine the spelling of the name appropriately. Success in multimodal unsupervised learning endeavors would extend the lifetime of deployed systems, and directly advance our ability to develop speech systems in new languages and domains without onerous demands of expensive human-labeled data, essentially by creating systems that automatically adapt and improve over time.

Portability and generalizability. An important aspect of learning is generalization. When a small amount of test data is available to adjust speech recognizers, we call such generalization adaptation. Adaptation and generalization capabilities enable rapid speech recognition application integration. There are also attempts to use partially observable Markov decision processes to improve dialogue management if training data can be made available.[42] This set of language resources is often not readily available for many new languages or new tasks. Indeed, obtaining large quantities of training data that is closely matched to the domain is perhaps the single most reliable method to make speech systems work in practice.

Over the past three decades, the speech community has developed and refined an experimental methodology that has helped to foster steady improvements in speech technology. The approach that has worked well is to develop shared corpora, software tools, and guidelines that can be used to reduce differences between experimental setups down to the algorithms, so it becomes easier to quantify fundamental improvements. Typically, these corpora are focused on a particular task. Unfortunately, current language models are not easily portable across different tasks as they lack linguistic sophistication to consistently distinguish meaningful sentences from meaningless ones. Discourse structure is not considered either, merely the local collocation of words.

This strategy is quite different from the human experience. For our entire lives, we are exposed to all kinds of speech data from uncontrolled environments, speakers, and topics, (that is, everyday speech). Despite this variation in our own personal training data we are all able to create internal models of speech and language that are remarkably adept at dealing with variation in the speech chain. This ability to generalize is a key aspect of human speech processing that has not yet found its way into modern speech systems. Research activities on this topic should produce technology that will operate more effectively in novel circumstances, and that can generalize better from smaller amounts of data. Another research area could explore how well information gleaned from large resource languages and/or domains generalize to smaller resource languages and domains.

The challenge here is to create spoken language technologies that are rapidly portable. To prepare for rapid development of such spoken language systems, a new paradigm is needed to study speech and acoustic units that are more language-universal than language-specific phones. Three specific research issues must be addressed: cross-language acoustic modeling of speech and acoustic units for a new target language; cross-lingual lexical modeling of word pronunciations for new language; and cross-lingual language modeling. By exploring correlation between new languages and well-studied languages, we can facilitate rapid portability and generalization. Bootstrapping techniques are keys to building preliminary systems from a small amount of labeled utterances, using them to label more utterance examples in an unsupervised manner, and iterating to improve the systems until they reach a comparable performance level similar to today's high-accuracy systems.

Dealing with uncertainties. The proven statistical DNN-HMM learning framework requires massive amounts of data to deal with uncertainties. How to identify and handle a multitude of variability factors has been key to building successful speech recognition systems. Despite the impressive progress over the past decades, today's speech recognition systems still degrade catastrophically even when the deviations are small in the sense the human listener exhibits little or no difficulty. Robustness of speech recognition remains a major research challenge. We hope for breakthroughs not only in algorithms but also in using the increasingly unsupervised training data available in ways not feasible before.

One pervasive type of variability in the speech signal is the acoustic environment. This includes background noise, room reverberation, the channel through which the speech is acquired (such as cellular, Bluetooth, landline, and VoIP), overlapping speech, and Lombard or hyper-articulated speech. The acoustic environment in which the speech is captured and the communication channel through which the speech signal is transmitted represent significant causes of harmful variability that is responsible for drastic degradation of system performance. Existing techniques are able to reduce variability caused by additive noise or linear distortions, as well as compensate for slowly varying linear channels. However, more complex channel distortions such as reverberation or fast-changing noise, as well as the Lombard effect present a significant challenge. While deep learning enabled auto-encoding to create more powerful features, we expect more breakthroughs in learning useful features that may or may not resemble imitating human auditory systems.

For the most part, algorithm designers for speech systems have ignored investigation of parallelism, partly because the advance of scalability has been so reliable.

Another common type of speech variability studied intensively is due to different speakers' characteristics. It is well known that speech characteristics vary widely among speakers due to many factors, including speaker physiology, speaker style, and accents—both regional and non-native. The primary method currently used for making speech recognition systems more robust is to include a wide range of speakers (and speaking styles) in the training, so as to account for the variations in speaker characteristics. Further, current speech recognition systems assume a pronunciation lexicon that models native speakers of a language and train on large amounts of speech data from various native speakers of the language. Approaches have been explored in modeling accented speech, including explicit modeling of accented speech, adaptation of native acoustic models with only moderate success, as witnessed by some initial difficulties of deploying British English speech system in Scotland. Pronunciation variants have also been incorporated in the lexicon to receive only small gains. Similarly, small progress has been made for detecting speaking rate change.

Having Socrates' wisdom. Like most of the ancient Greeks, speech recognition systems lack the wisdom of Socrates. The challenge here is to create systems that reliably detect when they do not know a (correct) word. A clue to the occurrence of such error events is the mismatch between an analysis of a purely sensory signal unencumbered by prior knowledge, such as unconstrained phone recognition, and a word- or phrase-level hypothesis based on higher-level knowledge, often encoded in a language model. A key component of this research would be to develop novel confidence measures and accurate models of uncertainty based on the discrepancy between sensory evidence and a priori beliefs. A natural sequel to detection of such events would be to transcribe them phonetically when the system is confident that its word hypothesis is unreliable, and to devise error-correction schemes.

Current systems have difficulty in handling unexpected—and thus often the most information rich—lexical items. This is especially problematic in speech that contains interjections or foreign or out-of-vocabulary words, and in languages for which there is relatively little data with which to build the system's vocabulary and pronunciation lexicon. A common outcome in this situation is that high-value terms are overconfidently misrecognized as some other common and similar-sounding word. Yet, such spoken events are key to tasks such as spoken term detection and information extraction from speech. Their accurate detection is therefore of vital importance.

Back to Top

Conclusion

Over the last four decades, there have been a number of breakthroughs in speech recognition technologies that have led to the solution of previously impossible tasks. Here, we will summarize the insights gained from the research and product development advances.

In 1976, the computational power available was only adequate to perform speech recognition on highly constrained tasks with low branching factors (perplexity). Today, we are able to handle nearly unlimited vocabularies with much larger branching factors. In 1976, the fastest computer available for routine speech research was a dedicated PDP-10 with 4MB memory. Today's systems have access to a million times more computational power in training the model. Thousands of processors and nearly unlimited collective memory capacity in the cloud are routinely used. These systems can use millions of hours of speech data collected from millions of people from the open population. The power of these systems arises mainly from their ability to collect, process, and learn from very large datasets.

The basic learning and decoding algorithms have not changed substantially in 40 years. However, many algorithmic improvements have been made, such as how to use distributed algorithms for the deep learning task. Surprisingly, even though there is probably enough computational power and memory in iPhone-like smartphone devices, it appears that speech recognition is currently done on remote servers with the results being available within a few hundred milliseconds on the iPhone. This makes it difficult to dynamically adapt to the speaker and the environment, which have the potential to reduce the error rate by half.

Dealing with previously unknown words continues to be a problem for most systems. Collecting very large vocabularies based on Web-based profiling makes it likely that the user would almost always use one of the known words. Today's Web search engines store over 500 million entity entries, which can be powerful to augment the vocabulary that is typically much smaller for speech recognition. The social graph used for Web search engines can also be used to dramatically reduce the needed search space. One final point is that mixed-lingual speech, where phrases from two or more languages may be intermixed, makes the new word problem more difficult.[17] This is often the case for many countries where English is mixed with the native language.

The associated problem of error detection and correction leads to difficult user interface choices for which good enough solutions have been adopted by "Dragon NaturallySpeaking" and subsequent systems. We believe multimodal interactive metaphor will be a dominant metaphor as illustrated by MiPad demo[16] and Apple Siri-like services. We are still missing human-like clarification dialog for new words previously unknown to the system.

Another related problem is the recognition of highly confusable words. Such systems require the use of more powerful discrimination learning. Dynamic sparse data learning, as is routinely done by human beings, is also missing in most of the systems that depend on large data-based statistical techniques.

Speech recognition in the next 40 years will pass the Turing test. It will truly bring the vision of Star Trek-like mobile devices to reality. We expect speech recognition to help bridge the gap between us and machines. It will be a powerful

tool to facilitate and enhance natural conservation among people regardless of barriers of location or language, as the *New York Times* story[a] illustrated by Rick Rashid's English to Chinese speech translation demo.[b]

Back to Top

References

1. Bahl, L. et al. Maximum mutual information estimation of HMM parameters. In *Proceedings of ICASSP* (1986), 49–52.

2. Baker, J. Stochastic modeling for ASR. *Speech Recognition*. D.R. Reddy, ed. Academic Press, 1975.

3. Baum, L. Statistical Estimation for Probabilistic Functions of a Markov Process. *Inequalities* III, (1972), 1–8.

4. Chen, X., et al. Pipelined back-propagation for context-dependent deep neural networks. In *Proceedings of Interspeech*, 2012.

5. Dahl, G., et al. Context-dependent pre-trained deep neural networks for LVSR. In *IEEE Trans. ASLP 20*, 1 (2012), 30–42.

6. Davis, S. et al. Comparison of parametric representations. *IEEE Trans ASSP 28*, 4 (1980), 357–366.

7. Dean, J. et al. Large scale distributed deep networks. In *Proceedings of NIPS* (Lake Tahoe, NV, 2012).

8. Dempster, et al. Maximum likelihood from incomplete data via the EM algorithm. *JRSS 39*, 1 (1977), 1–38.

9. De Mori, R. Spoken Dialogue with Computers. Academic Press, 1998.

10. Deng, L. and Huang, X. (2004). Challenges in adopting speech recognition. *Commun. ACM 47*, 1 (Jan. 2004), 69–75.

11. Deng, L. et al. Binary coding of speech spectrograms using a deep auto-encoder. In *Proceedings of Interspeech*, 2010.

12. Fiscus, J. Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE ASRU Workshop* (1997), 347–354.

13. He, X., et al. Discriminative learning in sequential pattern recognition. *IEEE Signal Processing 25*, 5 (2008), 14–36.

14. Hinton, G., et al. Deep neural networks for acoustic modeling in SR. *IEEE Signal Processing 29*, 11 (2012).

15. Huang, X., Acero, A., and Hon, H. *Spoken Language Processing*. Prentice Hall, Upper Saddle River, NJ, 2001.

16. Huang, X. et al. MiPad: A multimodal interaction prototype. In *Proceedings of ICASSP* (Salt Lake City, UT, 2001).

17. Huang, J. et al. Cross-language knowledge transfer using multilingual DNN. In *Proceedings of ICASSP* (2013), 7304–7308.

18. Hwang, M., and Huang, X. Shared-distribution HMMs for speech. *IEEE Trans S&AP 1*, 4 (1993), 414–420.

19. Jelinek, F. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1997.

20. Jelinek, F. Continuous speech recognition by statistical methods. In *Proceedings of the IEEE 64*, 4 (1976), 532–557.

21. Katagiri, S. et al. Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. In *Proceedings of the IEEE 86*, 11 (1998), 2345–2373.

22. Kingsbury, B. et al. Scalable minimum Bayes risk training of deep neural network acoustic models. In *Proceedings of Interspeech* 2012.

23. Klatt, D.H. Review of the ARPA speech understanding project. *JASA 62*, 6 (1977), 1345–1366.

24. Lee, C. and Huo, Q. On adaptive decision rules and decision parameters adaption for ASR. In *Proceedings of the IEEE 88*, 8 (2000), 1241–1269.

25. Lee, K. *ASR: The Development of the Sphinx Recognition System*. Springer-Verlag, 1988.

26. Lowerre, B. The Harpy Speech Recognition System. Ph.D. Thesis (1976). Carnegie Mellon University.

27. Mikolov, T. et al. Extensions of recurrent neural network language model. In *Proceedings of ICASSP* (2011), 5528–5531.

28. Mohri, M. et al. Weighted finite state transducers in speech recognition. *Computer Speech & Language 16* (2002), 69–88.

29. Morgan, N. et al. Continuous speech recognition using mulitlayer perceptions with Hidden Markov Models. In *Proceedings of ICASSP* (1990).

30. Pieraccini R. et al. A speech understanding system based on statistical representation. In *Proceedings of ICASSP* (1992), 193–196.

31. Potter, R., Kopp, G. and Green, H. *Visible Speech*. Van Nostrand, New York, NY, 1947.

32. Price, P. Evaluation of spoken language systems: The ATIS domain. In *Proceedings of the DARPA Workshop*, (Hidden Valley, PA, 1990).

33. Rabiner L. and Juang, B. *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

34. Reddy, R. Speech recognition by machine: A review. In *Proceedings of the IEEE 64*, 4 (1976), 501–531; http://www.rr.cs.cmu.edu/sr.pdf.

35. Seneff S. Tina: A NL system for spoken language application. *Computational Linguistics 18*, 1 (1992), 61–86.

36. Tur, G., and De Mori, R. *SLU: Systems for Extracting Semantic Information from Speech*. Wiley, U.K., 2011.

37. Yan, Z., Huo, Q., and Xu, J. A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR. In *Proceedings of Interspeech* (2013).

38. Yao, K. et al. Recurrent neural networks for language understanding. In *Proceedings of Interspeech* (2013), 104–108.

39. Yu, D. et al. Feature learning in DNN—Studies on speech recognition tasks. *ICLR* (2013).

40. Waibel, A. Phone recognition using time-delay neural networks. *IEEE Trans. on ASSP 37*, 3 (1989), 328–339.

41. Ward, W. et al. Recent improvements in the CMU SUS. In *Proceedings of ARPA Human Language Technology* (1994), 213–216.

42. Williams, J. and Young, S. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language 21*, 2 (2007), 393–422.

43. Zue, V. The use of speech knowledge in speech recognition. In *Proceedings of the IEEE 73*, 11 (1985), 1602–1615.

Back to Top

Authors

Xuedong Huang is a Distinguished Engineer of Bing Core Search at Microsoft Corp., Redmond, WA, where he founded its Speech Technology Group in 1993. He was previously on the faculty of Carnegie Mellon Univerisity.

James Baker is a former chair, CEO, and co-founder of Dragon Systems in Newton, MA. He received his Ph.D. from Carnegie Mellon University.

Raj Reddy is the Moza Bint Nasser University Professor of Computer Science and Robotics at Carnegie Mellon University in Pittsburgh, PA. He joined CMU in 1969.

Back to Top

Footnotes

a. http://nyti.ms/190won1

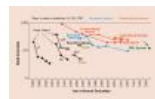b. https://www.youtube.com/watch?v=Nu-nlQqFCKg

Back to Top

Figures

Figure 1. Historical progress of speech recognition word error rate on more and more difficult tasks.[10] The latest system for the switchboard task is marked with the green dot.

Figure 2. Modern search engines such as Bing and Google both offer a readily accessible microphone button (marked in red) to enable voice search the Web. Apple iPhone Siri, while not a search engine (its Web search is now powered by Bing), has a much larger microphone button for multimodal speech dialogue.

Figure 3. There is no data like more data. Recognition word error rate vs. the amount of training hours for illustrative purposes only. This figure illustrates how modern speech recognition systems can benefit from increased training data.

Back to Top

Sidebar: What we did not know how to do in 1976.v

Statistical modeling and machine learning: Elaboration of HMM, context-dependent phoneme modeling, statistical smoothing and back-off strategies, DNN, semi-supervised learning, discriminative training such as Maximum Mutual Information Estimation (MMIE) and MPE

Training data and computing resources: Several orders of magnitude increase in the size of speech (thousands of hours) and text data (trillions of words) accompanied by the steadily increased distributed CPU and RAM resources

Signal processing dealing with noisy environments: DNN-learned features, MFCC appropriate for Gaussian mixture models, lower-level raw features such as filterbanks appropriate for DNN, Cepstral mean subtraction, 1st and 2nd order delta features, online environment adaptation, and noise-canceling microphone/microphone array

Vocabulary size and dis-fluent speech: From thousands to millions of words supported by n-grams and RNN as the language model, explicit garbage models, and the flexibility to add new words with grapheme form

Speaker independent and adaptive speech recognition: Mixture distributions, speaker training data across different dialects and populations, vocal tract normalization, Maximum a Posteriori (MAP), Maximum Likelihood Linear

Regression (MLLR), and unsupervised speaker-adaptive learning

Efficient decoder: Time-synchronous Viterbi search and A* stack decoder with sophisticated pruning techniques, distributed implementation to support large-scale server-based runtime decoder

Spoken language understanding and dialog: Case-frame based robust parser, semi-Markov conditional random field (CRF), boosted decision tree, rule-based or Markov decision process-based dialog management, and recurrent neural networks for sentence understanding

---

No entries found

Comment on this article

Signed comments submitted to this site are moderated and will appear if they are relevant to the topic and not abusive. Your comment will appear with your username if published. View our policy on comments

(Please sign in or create an ACM Web Account to access this feature.)

Create an Account

SUBMIT FOR REVIEW