Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 60 (2013) 16239-16243

Awakening to reality

Elixir
ISSN: 2229-712X

Evolution of Speech Recognition – A Brief History of Technology Development

Suma Swamy[1] and K.V Ramakrishnan[2]

[1]Department of Computer Science & Engineering, Sir M Visvesvaraya Institute of Technology, Bangalore.
[2]Department of Electronics & Communication Engineering, Anna University of Technology, Chennai.

ARTICLE INFO

Article history:
Received: 9 December 2012;
Received in revised form:
24 June 2013;
Accepted: 6 July 2013;

Keywords
ASR,
PAT,
OVE,
DARPA,
HMM.

ABSTRACT

This paper gives a brief overview of evolution of speech recognition starting from 1779 till date. It also discuss about the past, present and future of speech recognition.

Introduction

Speech Recognition, also known as automatic speech recognition (ASR), computer speech recognition, or erroneously as voice recognition, is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program, or more simply, as the ability of machines to respond to spoken commands[1].

Early speech recognition tried to apply grammatical and syntactical rules to speech, and when the words would fit into a certain set of rules the software was then able to figure out what was trying to be said. But the human language has much variation that made this type of software not have a very high accuracy level. Today speech recongition systems use statistical modeling systems, which use a mathmatical functions and probability [1].

Speech Technology

Three primary speech technologies are used in voice processing applications: stored speech, text-to- speech and speech recognition. Stored speech involves the production of computer speech from an actual human voice that is stored in a computer's memory and used in any of several ways [4].

Speech can also be synthesized from plain text in a process known as text-to-speech which also enables voice processing applications to read from textual database. Speech recognition is the process of deriving either a textual transcription or some form of meaning from a spoken input. Speech analysis can be thought of as that part of voice processing that converts human speech to digital forms suitable for transmission or storage by computers. Speech synthesis functions are essentially the inverse of speech analysis – they reconvert speech data from a digital form to one that's similar to the original recording and suitable for playback. Speech analysis processes can also be referred to as a digital speech encoding (or simply coding) and speech synthesis can be referred to as Speech decoding [4].

II History of Speech Recognition

Speech recognition research has been on going for more than 80 years. Over that period there have been at least 4 generations of approaches, and we forecast a 5th generation that is being formulated based on current research themes. The 5 generations, and the technology themes associated with each of them, are as follows [1].

A. Early Efforts - Mechanical Synthesis

The earliest efforts to produce synthetic speech began over two hundred years ago. In St. Petersburg in 1779, Professor Christian Kratzenstein explained the physiological differences between five long vowels and made an apparatus to produce them artificially. He constructed acoustic resonators similar to the human vocal tract and activated the resonators with vibrating reeds like in music instruments [1].

In 1791, in Vienna, Wolfgang von Kempelen introduced an "Acoustic-Mechanical Speech Machine" which could produce single sounds and some combinations of sounds. His machine had a pressure chamber for the lungs, a vibrating reed which acted as vocal cords, and a leather tube for vocal tract action. By manipulating the shape of the tube, he could produce different vowel sounds. Kempelen received negative publicity and was not taken seriously due to other inventions of his that proved to be fraudulent. However, his machine did lead to new theories regarding human vocals [1].

In the mid 1800's, Charles Wheatstone constructed a version of von Kempelen's speaking machine which could produce vowels and most consonant sounds. He could even produce some full words. Alexander Graham Bell, inspired by Wheatstone's machine, also constructed a similar machine of his own [1].

Research and experiments with mechanical vocal systems continued until the 1960's, but with little success. It was not until the beginning of electrical synthesizers that voice recognition began its true beginning [1].

Tele:
E-mail addresses: suma_swamy@yahoo.com

B. 1930's-'40s - Homer Dudley and Vocoder

Engineer Homer Dudley at AT&T's Bell Labs conducted the first experiments in voice encoding in 1928. At that time, the labs produced the first electronic speech synthesizer called a machine called Voder or Vocoder, derived from voice encoder. Dudley patented his invention in 1935. He, along with fellow engineers Riesz and Watkins produced the first electronic speech synthesizer in 1936. Experts who used a keyboard demonstrated it in the 1939 Worlds Fairs and foot pedals to play the machine and emit speech. The Vocoder was originally developed as a speech coder for telecommunications applications in the 1930s, with the idea being to code speech for transmission. In this fashion, it was used for secure radio communication, where voice was digitized, encrypted, and then transmitted on a narrow, voice-bandwidth channel [1].

Most early research in voice encoding and speech recognition was funded and performed by Universities and the U.S. Government, primarily by the military and the Defense Advanced Research Project Agency. Dudley's Vocoder was used in the SIGSALY system, built by Bell Labs engineers in 1943. This system was used for encrypting high-level communications for Allies during World War II. After the '30s, and early '40s, there was little improvement on Dudley's Vocoder and speech recognition [1].

C. 1950's - '60s – Synthesizers

In 1951, Franklin Cooper developed a Pattern Playback synthesizer at Haskins Laboratories. It reconverted recorded spectrogram patterns into sounds, either in original or modified forms. They were recorded optically on a transparent belt. Spectrograms calculate the frequency spectrum of a compound signal. It is a three-dimensional plot of the energy of the frequency content of a signal as it changes over time [1].

In 1953, Walter Lawrence introduced the first formant synthesizer, PAT (Parametric Artificial Talker). It consisted of three electronic formant resonators connected in parallel. A buzz or noise was inputted and a moving glass slide converted painted patterns into six time functions to control the three formant frequencies: voicing amplitude, fundamental frequency, and noise amplitude. It was the first successful synthesizer to describe the reconstruction process in terms of vocal tract resonances [1].

At the same time, Gunner Fant introduced the first cascade formant synthesizer called OVE I (Orator Verbis Electris), which consisted of resonators, connected in cascade. In 1962, Fant and his colleague Martony introduced the OVE II synthesizer, which had separate parts to model the transfer function of the vocal tract for vowels, nasals, and consonants. These synthesizers lead to the OVE III and GLOVE projects [1].
Baby Talk: The first speech recognition systems could understand only digits. (Given the complexity of human language, it makes sense that inventors and engineers first focused on numbers.) Bell Laboratories designed in 1952 the "Audrey" system, which recognized digits spoken by a single voice. Ten years later, IBM demonstrated at the 1962 World's Fair its "Shoebox" machine, which could understand 16 words spoken in English.

Labs in the United States, Japan, England, and the Soviet Union developed other hardware dedicated to recognizing spoken sounds, expanding speech recognition technology to support four vowels and nine consonants.

They may not sound like much, but these first efforts were an impressive start, especially when you consider how primitive computers themselves were at that time.[6]

D. 1970's to 2001 - The HMM Model and the Commercial Market

In the early 1970's, Lenny Baum of Princeton University invented the Hidden Markov Modeling approach to speech recognition. This is a statistical model which outputs a sequence of symbols or quantities and matches patterns. This approach became the basis for modern speech recognition and was adopted by all leading speech recognition companies. Baum shared his invention with several Advanced Research Projects Agency contractors including IBM [1].

In 1971, DARPA (Defense Advanced Research Projects Agency) established the Speech Understanding Research program to develop a computer system that could understand continuous speech. Lawrence Roberts spent $3 million per year for five years of government funds on the program. This led to the establishment of many Speech Understanding Research groups and was the largest speech recognition project ever [1].
Speech Recognition Takes Off: Speech recognition technology made major strides in the 1970s, thanks to interest and funding from the U.S. Department of Defense. The DoD's DARPA Speech Understanding Research (SUR) program, from 1971 to 1976, was one of the largest of its kind in the history of speech recognition, and among other things it was responsible for Carnegie Mellon's "Harpy" speech-understanding system. Harpy could understand 1011 words, approximately the vocabulary of an average three-year-old.

Harpy was significant because it introduced a more efficient search approach, called *beam search*, to "prove the finite-state network of possible sentences," according to Readings in Speech Recognition by Alex Waibel and Kai-Fu Lee. (The story of speech recognition is very much tied to advances in search methodology and technology, as Google's entrance into speech recognition on mobile devices proved just a few years ago.)

The '70s also marked a few other important milestones in speech recognition technology, including the founding of the first commercial speech recognition company, Threshold Technology, as well as Bell Laboratories' introduction of a system that could interpret multiple people's voices.[6]
In 1978, Texas Instruments introduced a popular toy called "Speak and Spell". It used a speech chip which led to huge strides in development of more human like digital synthesis sound [1].

1980s Speech Recognition Turns toward Prediction: Over the next decade, thanks to new approaches to understanding what people say, speech recognition vocabulary jumped from about a few hundred words to several thousand words, and had the potential to recognize an unlimited number of words. One major reason was a new statistical method known as the hidden Markov model.

Rather than simply using templates for words and looking for sound patterns, HMM considered the *probability* of unknown sounds' being words. This foundation would be in place for the next two decades. [6]
In 1982, Dragon Systems was founded by doctors Jim and Janet Baker. It has a long history of speech and language technology innovations and patents. In 1984, SpeechWorks, which is the leading provider of over-the-phone automated speech recognition, was founded [1].

Equipped with this expanded vocabulary, speech recognition started to work its way into commercial applications for business and specialized industry (for instance, medical use). It even entered the home, in the form of Worlds of Wonder's Julie doll (1987), which children could train to respond to their voice. ("Finally, the doll that understands you.")

However, whether speech recognition software at the time could recognize 1000 words, as the 1985 Kurzweil text-to-speech program did, or whether it could support a 5000-word vocabulary, as IBM's system did, a significant hurdle remained: These programs took discrete dictation, so you had to pause after each and every word.[6]

1990s Automatic Speech Recognition Comes to the Masses: In the '90s, computers with faster processors finally arrived, and speech recognition software became viable for ordinary people.

In 1990, Dragon launched the first consumer speech recognition product, Dragon Dictate, for an incredible price of $9000. Seven years later, the much-improved Dragon NaturallySpeaking arrived. The application recognized continuous speech, so you could speak, well, naturally, at about 100 words per minute. However, you had to train the program for 45 minutes, and it was still expensive at $695.

The advent of the first voice portal, VAL from BellSouth, was in 1996; VAL was a dial-in interactive voice recognition system that was supposed to give you information based on what you said on the phone. VAL paved the way for all the inaccurate voice-activated menus that would plague callers for the next 15 years and beyond.[6]

n 1995, Dragon Systems released word dictation-level speech recognition software, which was the first time dictation speech recognition technology was available to consumers. IBM and Kurzweil soon followed the trend [1].

Charles Schwab became the first company to devote resources towards developing the program Voice Broker in 1996. The program allows for up to 360 simultaneous customers to call in and get quotes on stocks and options and handled 50,000 requests daily. It was 95% accurate and set the stage for many other companies to follow in their footsteps. BellSouth launched the first voice portal called Val, which is a type of web portal that can be accessed be people entirely by voice, used by both consumers and businesses [1].

In 1997, Dragon introduced "Naturally Speaking" which was the first "continuous speech" dictation software available, meaning that you no longer needed to pause between words for the computer to understand what was being said [1].

In 1998, Lernout and Hauspie bought Kurzweil. Microsoft invested $45 million in Lernout and Hauspie to form a partnership, eventually allowing Microsoft to use their speech recognition technology in their systems. In 1999, Microsoft acquired Entropic, giving them access to the "most accurate speech recognition system in the world"[1].

In 2000, Lernout and Hauspie acquired Dragon Systems for $460 million. In the same year, TellMe introduced the first world-wide voice portal [1].

In 2001, ScanSoft acquired Lernout and Hauspie as well as their speech and language assets. They also acquired SpeechWorks in 2003 as well as closing a deal to distribute and support IBM desktop products that employ speech recognition [1].

2000s Speech Recognition Plateaus--Until Google Comes Along: By 2001, computer speech recognition had topped out at 80 percent accuracy, and, near the end of the decade, the technology's progress seemed to be stalled. Recognition systems did well when the language universe was limited--but they were still "guessing," with the assistance of statistical models, among similar-sounding words, and the known language universe continued to grow as the Internet grew.

Speech recognition and voice commands were built into Windows Vista and Mac OS X. Many computer users weren't aware that those features existed. Windows Speech Recognition and OS X's voice commands were interesting, but not as accurate or as easy to use as a plain old keyboard and mouse.[6]

E. 2001 to 2010

Use of parallel processing methods to increase recognition decision reliability; combinations of HMMs and acoustic-phonetic approaches to detect and correct linguistic irregularities; increased robustness for recognition of speech in noise; machine learning of optimal combinations of models [1].

F. 2010 to 2013- Biometrics and market

Automatic Speech Recognition Market – 2009 -2013-report forecasts that the market will reach $933 million in 2013 growing at 15 percent.[5]

Speech recognition technology development began to edge back into the forefront with one major event: the arrival of the Google Voice Search app for the iPhone. The impact of Google's app is significant for two reasons. First, cell phones and other mobile devices are ideal vehicles for speech recognition, as the desire to replace their tiny on-screen keyboards serves as an incentive to develop better, alternative input methods. Second, Google had the ability to offload the processing for its app to its cloud data centers, harnessing all that computing power to perform the large-scale data analysis necessary to make matches between the user's words and the enormous number of human-speech examples it gathered.

In short, the bottleneck with speech recognition has always been the availability of data, and the ability to process it efficiently. Google's app adds, to its analysis, the data from billions of search queries, to better predict what you're probably saying.

In 2010, Google added "personalized recognition" to Voice Search on Android phones, so that the software could record users' voice searches and produce a more accurate speech model. The company also added Voice Search to its Chrome browser in mid-2011. Remember how we started with 10 to 100 words, and then graduated to a few thousand? Google's English Voice Search system now incorporates 230 billion words from actual user queries.

And now along comes Siri. Like Google's Voice Search, Siri relies on cloud-based processing. It draws what it knows about you to generate a contextual reply, and it responds to your voice input with personality.

Speech recognition has gone from utility to entertainment. The child seems all grown up.

The Future: Accurate, Ubiquitous Speech: The explosion of voice recognition apps indicates that speech recognition's time has come, and that you can expect plenty more apps in the future. These apps will not only let you control your PC by voice or convert voice to text--they'll also support multiple languages, offer assorted speaker voices for you to choose from, and integrate into every part of your mobile devices (that is, they'll overcome Siri's shortcomings).

The quality of speech recognition apps will improve, too. For instance, Sensory's Trulyhandsfree Voice Control can hear and understand you, even in noisy environments.

As everyone starts becoming more comfortable speaking aloud to their mobile gadgets, speech recognition technology will likely spill over into other types of devices. It isn't hard to imagine a near future when we'll be commanding our coffee makers, talking to our printers, and telling the lights to turn themselves off.[6]

Key Market Drivers include:

- Success stories of early adopters pushing peer companies that continue to use manual processes
- Growing use of biometrics for secure identification
- purposes

- Advancements in speech recognition technologies that has made it more accurate and robust
- Clear cost savings across industries with the use of speech recognition applications

Despite continuing to grow, this market has not reached the exponential growth predicted by earlier estimates. This can be attributed to apprehensive customers, unavailability of technology in some regions and high pricing pressure on mobile devices. High R&D expenses have also added to this. [5]

III Future OF SPEECH recognition

Speech recognition should reach a point where the computer is able to accurately transcribe close to 100% of what is said. This belief is backed by the principal of Moore's Law which anticipates a substantial increase in available memory, capacity, and computer power. The key is to teach the machines vocabulary phonetically so that it can understand what it is hearing and assign specific sounds to each combination of letters. Already, speech recognition has exceeded the productivity of manual labor in some areas. For example, when there is a large quantity of data, like there is in a directory, machines are already more capable of sifting through the great amount of data than people [1].

In the distant future, speech recognition may turn into speech understanding, meaning that the machines would not only recognize the words that are being said, but would also be able to understand what they mean. In fact, there are people who believe that eventually, the computers will have the capability of talking back, and even carrying on a conversation. However, even humans are imperfect, and the expectation cannot be that the machines will make no mistakes. A human scribe will sometimes mishear or mistype a sentence or two, and nothing more should be expected of these machines. They are merely assuming the responsibilities of a human assistant with an equal amount of productivity, in turn enabling human assistants to focus their time on something that machines have not been invented for yet [1].

The Defense Advanced Research Projects Agency is developing a program, Global Autonomous Language Exploitation that will translate international news broadcasts and newspapers. The goal is to produce a product that can translate two languages with at least 90% accuracy. The ultimate goal, however, is a universal translator. This translator would be able to translate any language, but this product is still quite far from completion. The problem is that it is very difficult to combine speech recognition with automatic translation. Inconsistencies between languages such as slang, dialects, accents and background noises are difficult for a machine to recognize [1].

IV CONCLUSIONS

Speech technology has come a long way from its primitive origins, and the technology is currently at a stage where a significant number of useful and profitable applications can be made. Still, it is important to be aware of the limitations of the current technology, both for designing good and user-friendly systems today, and for aiming to solve the challenges of tomorrow. Speech technology is not mono- disciplinary, and in order to be able to create a speech- enabled machine that can pass the Turing test, the collaborative efforts of scientists from many areas are needed. Among the immediate challenges are: how to design speech recognizers that are robust to speaker accents and dialects as well as to noise contamination, how to create speech synthesizers that are able to convey emotions and can adopt speaking styles appropriate for any given text, how to deal with multi-linguality, how to attack the problem of recognizing spontaneous speech, just to name a few. Many problems have been solved, but speech technology research will not lack challenges in the years to come [3]

ACKNOWLEDGMENT

We would like to thank Visvesvaraya Technological University, Belguam and Anna University, Chennai for giving permission to publish tis paper. We would like to express our gratitude to our trust Sri Krishnadevaraya Educational Trust and ir M. Visvesvaraya InstitSute of Technology, Bangalore.

Table I Past, Present and future of speech recognition

Year	**Speech Recognition**		
	Technology	**Recognition**	**Application**
1771	acoustic resonators	five long vowels	
1791	Acoustic-Mechanical Speech Machine	vowel sounds, single and combinations of sounds	
1800-1900	speaking machine	*Vowels, consonants, full words*	
1928-1943	*Vocoder*		*Secure radio transmission*
1951	Pattern Playback synthesizer		recorded spectrogram patterns into sounds
1953	first formant synthesizer		the reconstruction process in terms of vocal tract resonances
1971	Speech Understanding Research program	continuous speech	
1978	"Speak and Spell".chip		digital synthesis sound.
1982-1984	Dragon Systems SpeechWorks		over-the-phone automated speech recognition
1996	Voice Broker		360 simultaneous customers to call
1997	"Naturally Speaking" dictation software	continuous speech	
2001-2010	parallel processing		detect and correct linguistic irregularities
2010-2013			*Biometrics*

Our Sincere thanks to Dr. M.S. Indira, Principal, Sir M. Visvesvaraya Institute of Technology, Bangalore for her support..

Our sincere thanks to Professor Dilip.K.Sen, Head of the Department of CSE and all other staff members for their invaluable suggestions from time to time, support and encouragement.

References

[1] History of Speech Recognition, acamedics.

[2] Automatic Speech Recognition, ASR News, July 2005, Volume 16 No. 7, Market, Investment and Technical News of the Emerging Speech Recognition Industry.

[3] Speech Technology: Past, Present and Future, Torbjorn Svendsen, Telektronikk Vol. 2.2003.
[4] Speech Recognition Seminar, Department of ECE, PESIT
[5] Key Market Drives: Market Report, Market Publishers.
[6] Speech Recognition Through the Decades: How We Ended Up With Siri, Melanie Pinola, PCWorld.