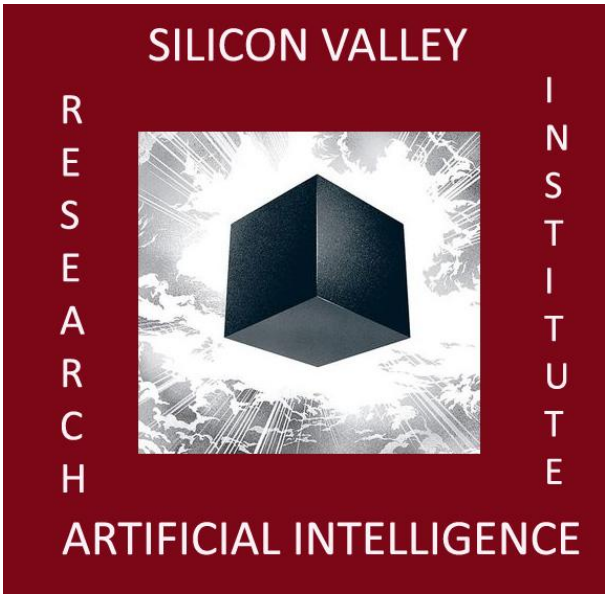

Speech Recognition: a brief introduction

A brief summary of the field of speech recognition can serve to explain the infinite number of practical problems that must be solved in order to have a machine simply understand the words that i am saying (never mind the meaning of those words, just the words). A vast gulf separates popular books on the Singularity from the mundane daily research carried out at A.I. laboratories, where scientists work on narrow specialized technical details. The history of speech recognition goes back at least to 1961, when IBM researchers developed the "Shoebox", a device that recognized spoken digits (0 to 9) and a handful of spoken words. In 1963 NEC of Japan developed a similar digit recognizer. Tom Martin at the RCA Laboratories was probably the first who applied neural networks to speech recognition ("Speech Recognition by Feature Abstraction Techniques", 1964). In 1970 Martin founded Threshold Technology in New Jersey which developed the first commercial speech-recognition product, the VIP-100.

Speech analysis became a viable technology thanks to conceptual innovations in Russia and Japan. In 1966 Fumitada Itakura at NTT in Tokyo invented Linear Predictive Coding ("One Consideration on Optimal Discrimination or Classification of Speech", 1966), a technique that 40 years later would be still used for voice compression in the GSM protocol for cellular phones; and Taras Vintsiuk at the Institute of Cybernetics in Kiev invented Dynamic Time Warping ("Speech Discrimination by Dynamic Programming", 1968), utilizing dynamic programming (a mathematical technique invented by Richard Bellman at RAND in 1953) to recognize words spoken at different speeds. Dynamic Time Warping was refined in 1970 by Hiroaki Sakoe and Seibi Chiba at NEC in Japan. Meanwhile, in 1969 Raj Reddy (who had been the first PhD graduate of the Stanford computer science department in 1966) founded the speech-recognition group at Carnegie Mellon University and supervised three important projects: Harpy (Bruce Lowerre 1976), that used a finite-state network to reduce the computational complexity; Hearsay-II, that pioneered the "blackboard" in which knowledge acquired by parallel

asynchronous processes gets integrated to produce higher levels of hypothesis, a blend of bottom-up and top-down processing (Rick Hayes-Roth, Lee Erman, Victor Lesser and Richard Fennell, 1975); and Dragon, developed in 1975 by Jim Baker, who then moved to Massachusetts to start a pioneering company with the same name in 1982. Dragon differed from Hearsay in the way it represented knowledge: Hearsay used the logical approach of the "expert system" school, whereas Dragon used the hidden Markov model. It was during the Hearsay project that Reddy invented the "beam search" algorithm to search large spaces of possible solutions.

Trivia: Dragon's technology was later acquired by Nuance, whose technology would be later acquired by a company named Siri that built a system for the Apple iPhone.

The same idea was central to Fred Jelinek's efforts at IBM ("Continuous Speech Recognition by Statistical Methods", 1976) and statistical methods based on the Hidden Markov Model for speech processing became popular with Jack Ferguson's "The Blue Book", which was the outcome of his 1980 lectures at the Institute for Defense Analyses in New Jersey.

IBM (Jelinek's group) and the Bell Labs (Lawrence Rabiner's group) came to represent two different schools of thought: IBM was looking for the individual speech-recognition system, that would be trained to recognize one specific voice; Bell Labs wanted a system that would understand a word pronounced by any one among the millions of AT&T's phone users. IBM studied the language model, whereas Bell Labs studied the acoustic model.

IBM's technology (the n-gram model) tried to optimize the recognition task by predicting statistically the next word. The inspiration for the IBM technique came from a word game devised by Claude Shannon in his book "A Mathematical Theory of Communication" (1948). Program this technique into a computer, and test it on your friends, and you have the Shannon equivalent of the Turing Test: ask both the computer and your friends to guess the next word in an arbitrary sentence. If the span of words is 1 or 2, your friends easily win. But if the span of words is 3 or higher, the computer starts winning.

Shannon's game was the first hint that perhaps understanding the meaning of the speech was irrelevant, and instead the frequency of each word and of its coexistence with other words was crucial.

Baum's Hidden Markov Model applied to speech recognition becomes a probability measure which integrates both schools because it can represent both the variability of speech sound and the structure of spoken language. The Bell Labs approach eventually led to Biing-Hwang Juang's "mixture-density hidden Markov models" for speaker independent recognition and a large vocabulary ("Maximum Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," 1985).

Hidden Markov Models became the backbone of the systems of the 1980s: Kai-Fu Lee's speaker-independent system Sphinx in 1988 at Carnegie Mellon University (the most successful system yet for a large vocabulary and continuous speech); the Byblos system from BBN (1989); and the Decipher system from SRI (1989).

Three projects further accelerated progress in speech recognition. In 1989 Steve Young at Cambridge University developed the Hidden Markov Model Tool Kit, which soon became the most popular tool to build speech-recognition software. During the 1990s at least two major speech recognition datasets were compiled, the CSR corpus and the Swtichboard corpus. Finally, in 1989 DARPA sponsored projects to develop speech recognition for air travel (the Air Travel Information Service or ATIS) with participants such as BBN, MIT, CMU, AT&T, SRI, etc. The program ended in 1994 when the yearly benchmark test showed that the error rate had dropped to human levels. These projects, largely based on Juang's algorithm of 1985, left behind another huge corpus of utterances. The following decade witnessed the first serious conversational agents: in 2000 Victor Zue at the MIT demonstrated Pegasus for airline flights status and Jupiter for weather status/forecast, and also in 2000 Al Gorin at AT&T developed How May I Help You (HMIHY) for telephone customer care. More importantly, the leader of the ATIS project at SRI, Michael Cohen, founded Nuance in 1994 that developed the system licensed by Siri to make the 2010 app for the Apple iPhone (and Cohen was hired by Google in 2004).

Voice Signal Technologies was founded by Dan Roth in 1995 in Boston and in 2002 it provided the first voice-dialing system (on a Samsung A500 phone), followed in 2003 by a name-dialing system (for the A610). Another startup of voice recognition was Israel's Advanced Recognition Technology, acquired by Nuance in 2005 as was Voice Signal in 2007.

After Alex Waibel's time-delay network (1989), combining HMM and neural nets became commonplace and led to the speech-recognition systems of Hinton at the University of Toronto ("Deep Belief Networks for Phone Recognition," 2009) and of Dong Yu at Microsoft ("Conversational Speech Transcription Using Context-Dependent Deep Neural Networks", 2011). HMMs still outperformed deep neural networks in speech recognition, especially for large vocabularies. LSTM neural nets began to be used in speech recognition after Alex Graves' experiments at the University of Toronto (" Speech Recognition with Deep Recurrent Neural Networks", 2013). As deep neural nets became more feasible and affordable, the demise of hidden Markov models became more appealing. HMMs are rather complicated to manipulate. End-to-end neural-network architectures began to look simpler and more elegant.

In 2014 Alex Graves trained an LSTM neural network with his own method of connectionist temporal classification (CTC) of 2006 and obtained a speech recognition system that didn't have any HMM, but its error rate was higher than the error rate of HMM-based models, especially with homophones, words that sound alike ("Towards End-to-End Speech Recognition with Recurrent Neural Networks", 2014). In such hybrid systems of HMMs and deep neural networks, the temporal reasoning takes place within the HMM rather than the neural network. CTC training of neural networks forces the network to carry out that job. At the end of 2014 the (cumbersome) HMM component of a speech-recognition system was definitely dropped by Andrew Ng's group at Stanford. His system used a "language model" and "prefix beam search" to search through the space of possible word sequences. For example, given the prefix "Somebody stole his", a language model might indicate that the word "wallet" has a 50% chance of being the next word, and "phone" a 25% chance, and so on. This trick helped Ng to simplify the architecture and use a regular recurrent neural network instead of an LSTM neural network ("First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs", 2014). The word error rate was reduced dramatically. This architecture was employed in 2015 by Andrew

Ng's team (now at Baidu's labs in Silicon Valley) to build Deep Speech 2 ("End-to-End Speech Recognition in English and Mandarin", 2015), which, for the record, is composed of eleven layers (three layers of convolution, seven recurrent layers, and a fully-connected output layer trained with batch normalization.

In 2016 Microsoft, by replacing Dong Yu's hybrid system with three kinds of convolutional nets for acoustic modeling and an LSTM for language modeling, achieved human parity (it transcribed speech as well as a professional transcriptionist) on the NIST 2000 dataset. The three convolutional nets were a variant of VGG-16, a variant of ResNet and a variant called LACE (layer-wise context expansion with attention) of the time-delay neural network pioneered by Alex Waibel in 1989.

Speech recognition systems are becoming ubiquitous: Apple's Siri (2011), Google's Now (2012), Microsoft's Cortana (2013), Wit.ai (founded in 2013 by Alexandre Lebrun in Silicon Valley and acquired by Facebook in 2015), Amazon's Alexa (2014), Baidu's Deep Speech 2 (2015, developed in Silicon Valley, the foundation of Xiaoyu that was introduced in 2017), SoundHound's Hound (launched in 2016 by a Silicon Valley startup founded in 2004 by Keyvan Mohajer), etc.

The year 2018 was the year of the full-duplex chatbot, the chatbot can talk and listen at the same time. First came Microsoft's full-duplex version of its Xiaoice (developed by Li Zhou's team in China), then Google's Duplex (developed by Yaniv Leviathan's team). Microsoft then acquired Semantic Machines, a startup founded in 2014 in Berkeley by veterans such as Dan Roth, Apple Siri's chief scientist Larry Gillick (formerly at Dragon and Voice Signal), Dan Klein of UC Berkeley and Percy Liang of Stanford, as well as Klein's student David Hall, who in 2010 built the Overmind agent that beat a human master at the videogame StarCraft.

But these systems share one limitation: they are designed to work in controlled environments using clear speech. The limitations of today's speech recognition become obvious when you talk to the machine in a noisy context. Unfortunately, that is increasingly the natural context. We are packing more people in the confined spaces of cities. Therefore, most verbal interactions happen in the cacophony of the city: multiple conversations happening at a party, beeping devices around the speakers, traffic noise all around, maybe a television screen blaring a soccer game, barking dogs, the clatter of drinking and eating, a machine alarm, an ambulance siren. Not only is there background noise, but it is totally unpredictable. Humans still manage to understand each other in these noisy environments because they are naturally able to discriminate what is voice and what is not, and to recognize the voice of their friend (even when it is not the loudest sound in the room). In real-world work situations the use of voice commands can be counterproductive. The problem is not easy to solve. There are tools to reduce and even eliminate noise, echo and reverb, but the result of these operations is to weaken the very voice that the device is trying to understand. Then identifying individual speakers becomes harder. At the end of 2017 an unknown entity posted an "ideation challenge" on the Innocentive website offering a monetary reward for ideas precisely on how to tackle this problem.