Spoken Language Communication with Machines: The Long and Winding Road from Research to Business

Roberto Pieraccini and David Lubensky

IBM T.J. Watson Research Center, 1101 Kitchawan Road,
Route 134, Yorktown Heights, NY 10598, USA
{rpieracc, davidlu}@us.ibm.com

Abstract. This paper traces the history of spoken language communication with computers, from the first attempts in the 1950s, through the establishment of the theoretical foundations in the 1980s, to the incremental improvement phase of the 1990s and 2000s. Then a perspective is given on the current conversational technology market and industry, with an analysis of its business value and commercial models.

1 Introduction

One of the first speech recognition systems was built in 1952 by three AT&T Bell Laboratories scientists [1].The system could recognize sequences of digits spoken with pauses between them. The pioneers of automatic speech recognition (ASR) reported that […] *an accuracy varying between 97 and 99 percent is obtained when a single male speaker repeats any random series of digits.* However, the system required to be adjusted for each talker […] *if no adjustment is made, accuracy may fall to as low as 50 or 60 percent in a random digit series.* The Automatic Digit Recognition machine, dubbed Audrey, was completely built with analog electronic circuits and, although voice dialing was a much attractive solution for AT&T towards cost reduction in the long distance call business, it was never deployed commercially.

It took more than three decades for the algorithms and the technology of speech recognition to find a stable setting within the framework of statistical modeling. And it took two more decades of incremental improvement to reach an acceptable level of performance.

Only towards the beginning of this century, nearly fifty years after Audrey, did we witness the emergence of a fairly mature market and of a structured industry around conversational applications of the *computer-speech* technology. The principles on which modern speech recognition components operate are not dissimilar to those introduced in the late 1970s and early 1980s. Faster and cheaper computers and the availability of large amounts of transcribed speech data allowed a relentless incremental improvement in speech recognition accuracy. Even though automatic speech recognition performance is not perfect, it offers tremendous business benefits in many different applications.

There are many applications of speech recognition technology, ranging from dictation of reports on a desktop computer, to transcription of conversations between

M. Ali and F. Esposito (Eds.): IEA/AIE 2005, LNAI 3533, pp. 6 – 15, 2005.

agents and callers, to speech-to-speech translation. Although many applications are commercially exploited in different niches of the market, the industry is mostly evolving around *conversational systems*, aimed at customer self-service in call centers, or providing effective control of devices in automotive or mobile environments.

2 A Brief History of Automatic Speech Recognition Research

The early history of speech recognition technology is characterized by the so-called linguistic approach. Linguists describe the speech communication chain with several levels of competence: acoustic, phonetic, lexical, syntactic, semantic, and pragmatic. Although the precise mechanism that grants humans, among all animals, the mastering of a sophisticated language was, and still is, mostly inscrutable, there was a fairly general agreement in the scientific community that, for building speech understanding machines, *that mechanism* should be replicated. Thus, most of the approaches to the recognition of speech in the 1960s were based on the assumption that the speech signal needs to be first segmented into the constituent phonetic units. Sequences of phonetic units can then be grouped into words, words into phrases and syntactic constituents, and eventually one can reach a semantic interpretation of the message.

Although an obvious solution, the linguistic approach never produced satisfactory results since the acoustic variability of speech prevented the accurate segmentation of utterances into phonetic units. The phonetic variability of words, mainly due to coarticulation phenomena at the word junctures and speaker variability, concurrently with errorful strings of decoded phonetic hypotheses, caused errors in the lexical transcriptions, which propagated to erroneous syntactic and semantic interpretations.

At the end of the 1960s, practical usability of speech recognition was extremely doubtful, and the seriousness of the field was severely mined by the lack of practical results. That prompted John Pierce, executive vice-president of Bell Laboratories, renowned scientist and visionary in the field of satellite communications, to launch into a contentious attack to the whole field in an infamous letter to the Journal of the Acoustical Society of America [2].

Although Pierce's letter banned speech recognition research at Bell Laboratories for almost a decade, it did not arrest the enthusiasm of a few visionaries who saw the potential of the technology. ARPA[1], the Advance Research Project Agency of the US Department of Defense, against the opinion of an advisory committee headed by Pierce, started in 1971 a $15 million 5 year research program that went under the name of SUR: Speech Understanding Research. At the end of the program, in 1976, four systems [3] were evaluated, but unfortunately none of them matched the initial requirements of the program—less that 10% understanding error with a 1000 words vocabulary in near real time. Three of the systems were based on variations of classical AI rule-based inference applied to the linguistic approach. One of them was built by SDC, and the other two—HWIM and Hearsay II—were developed by BBN and

[1] The name of the research agency changed through the years. It was ARPA at its inception in 1958; it became DARPA—D as in Defense—in 1972. President Bill Clinton changed its name back to ARPA in 1993. The initial D was added again in 1996. Today, in year 2005, it is still called DARPA.

CMU respectively. The fourth system, Harpy [4], built by CMU, went very close to match the program requirements. In fact, Harpy was capable of understanding 95% of the evaluation sentences, but its real time performance was quite poor: 80 times real time—a 3 second sentence required 4 minutes of processing on a 0.4 MIPS PDP-KA 10. Rather than using classical AI inference, Harpy was based on a network of 15,000 interconnected nodes that represented all the possible utterances within the domain, with all the phonetic, lexical, and syntactic variations. The decoding was implemented as a *beam-search* variation of the dynamic programming algorithm [5] first published by Bellman in 1957.

Harpy was not the first application of dynamic programming to the problem of speech recognition. A dynamic programming algorithm for matching utterances to stored templates was experimented first by a Russian scientist, Vintsyuk [6], in 1968, and then by Sakoe and Chiba [7] in Japan in 1971. However, in the hype of the AI years of the 1970s, the dynamic programming—or Dynamic Time Warping (DTW)—approach was considered a *mere engineering trick* which was limited to the recognition of few words at the expense of a large amount of computation. It was a brute-force approach which did not have the elegance, and the *intelligence* of systems based on rule inference. However, DTW was easier to implement—it did not require linguistic expertise—and its performance on well defined speech recognition tasks was generally superior to the more complex rule based systems.

In the mid 1970s Fred Jelinek and his colleagues at IBM Research started working on a rigorous mathematical formulation of the speech recognition problem [8] based on fundamental work on stochastic processes carried out a few years earlier by Baum at IDA [9]. The IBM approach was based on statistical models of speech—Hidden Markov Models, or HMM—and word n-grams. The enormous advantage of the HMM approach as compared with all the other methods is in the possibility of learning automatically from virtually unlimited amounts data. However, it was not until the early 1980s, thanks to the experimental work and tutorial paper [10] by Larry Rabiner and his colleagues at Bell Laboratories, that the HMM approach became mainstream in virtually all speech research institutions around the world.

3 The Power of Evaluation

In the 1980s and 1990s speech recognition technology went through an incremental improvement phase. Although alternative techniques, such as artificial neural networks, were investigated, HMM and word n-grams remained the undisputed performers. However, towards the end of the 1980s, a general disbelief about the actual applicability of speech recognition to large vocabulary human-machine spoken language dialog was still present in the speech recognition research community.

Automatic dictation systems of the late 1980's and a few other industrial hands-eyes busy applications were the only demonstrable products of speech recognition technology. In 1988, a company called Dragon, founded by Jim and Janet Baker, former IBM researchers, demonstrated the first PC-based, 8,000 words speech recognition dictation system. The system was commercialized in 1990, but did not find much appeal in the market; the computational limitations still required users to speak with pauses between words. Although next generation of dictation products did not require a user to pause between words, the market for this application never really took off, and still remains a niche market[2].

In the second half of the 1980s most of the research centers started following the HMM/n-gram paradigm. However the efforts remained fragmented and it was difficult to measure progress. Around this time DARPA funded a new effort [11] focused at improving speech recognition performance. The major difference from the previous program, the 1971 SUR, was in the new focus that DARPA put in the on-going evaluation of speech recognition systems from the program participants. A common task with a fixed vocabulary, associated with shared training and test corpora, guaranteed the scientific rigorousness of the speech recognition performance assessment, which was administered through regular yearly evaluations by NIST, the National Institute of Standards and Technology (NIST). The new program, called Resource Management, was characterized by a corpus of read sentences belonging to a finite state grammar with a 1000 word vocabulary. Word accuracy, a well defined standard metric, was used for assessing the systems and measuring progress. A controlled, objective, and common evaluation paradigm had the effect of pushing the incremental improvement of speech recognition technology. In a few years, program participants pushed the word error rate from 10% to a few percent.

The Resource Management task was the first in a series of programs sponsored by DARPA with increasingly more complex and ambitious goals. Airline Travel Information Systems (ATIS) [12] followed in the early 1990s, and was focused on spoken language understanding. The common evaluation corpus in the ATIS project included *spontaneous queries* to a commercial flight database, whereas Resource Management sentences were *read* and defined by a fixed finite state grammar. ATIS forced the research community to realize, for the first time, the difficulties of spontaneous speech. Spontaneous speech is not grammatical, with a lot of disfluencies, such as repetitions, false starts, self corrections, and filled pauses. Here is an example of a spontaneous sentence from the ATIS corpus:

From um sss from the Philadelphia airport um at ooh the airline is United Airlines and it is flight number one ninety four once that one lands I need ground transportation to uh Broad street in Phileld Philadelphia what can you arrange for that.[3]

With ATIS, the speech recognition and understanding community realized that classical natural language parsing based on formal context free or higher order grammars failed on spontaneous speech. Statistical models, again, demonstrated their superiority in handling the idiosyncrasies of spoken language. A new paradigm invented at AT&T Bell Laboratories [13] and aimed at the detection of semantically meaningful phrases was soon adopted, in different forms, by several other institutions and demonstrated superior performance when compared with traditional parsing methods.

The ATIS program, which ended in 1994, while fostering the incremental improvement in the speech recognition and understanding performance, did not provide

[2] Besides providing accessibility to disabled individuals, speech recognition based dictation found most of the adopters within the professional community of radiologists.

[3] This sentence was judged to be the most ungrammatical spontaneous sentence among those recorded by the MADCOW committee in the early 1990s , during the DARPA ATIS project. A t-shirt with this sentence imprinted on the back was made available to program participants.

a satisfactory answer to the general problem of human-machine spoken communication. ATIS dialogs were mostly *user initiated*: the machine was only intended to provide answers to questions posed by the user. This is not a typical situation of regular conversations, where both parties can ask questions, provide answers, and change the course of the dialog. This situation, known as *mixed-initiative* dialog, contrasts with the other extreme case, where the machine asks questions and the user can only provide answers, known as *system-initiative*, or directed-dialog.

Aware of the important role of mixed initiative dialog systems in human-machine communication, DARPA followed the ATIS program with the launch of another project known as the DARPA Communicator [14]. Other programs with complex speech recognition tasks based on corpora, such as *Switchboard* (human-human conversational speech) and *Broadcast News* (broadcast speech) followed, until the most recent EARS[4] (Effective Affordable Reusable Speech-to-text). On-going DARPA evaluations push researchers to invent new algorithms and focus not only on speech recognition accuracy but also computational efficiency.

4 The Change of Perspective and the Conversational Market

While in the mid 1990s the research community was improving the speech recognition performance on more and more complex tasks, two small startup companies appeared on the quite empty market landscape. The first was Corona, renamed successively Nuance, a spin-off from the Stanford Research Institute (SRI). The second, an MIT spin-off, was initially called Altech, and then renamed SpeechWorks[5]. While the research community was focusing on complex human-like conversational dialog systems, SpeechWorks and Nuance took a different perspective. If the task is simple enough for the available speech recognition technology to attain reasonable accuracy, the interface can be engineered in such a way as to provide an excellent user experience, certainly superior to that offered by conventional touch-tone Interactive Voice Response (IVR). Simple applications, such as package tracking, where the user is only required to speak an alphanumeric sequence (with a checksum digit), and slightly more complex applications such as stock quote and flight information were their commercial targets.

The notion of user-centered design, as opposed to technology driven, became the guiding principle. Users want to accomplish their task with the minimal effort. Whether they can talk to machines with the same freedom of expression offered in human-human conversations, or they are gracefully directed to provide the required information in the simplest way, proved to be of little concern to users. Getting to the end of the transaction in the shortest time is the most important goal. Notwithstanding the efforts of the research community, free-form natural-language speech was, in the mid 1990s, still highly error prone. On the other hand, limiting the response of users to well crafted grammars provided enough accuracy to attain high levels of automation. In a way, SpeechWorks and Nuance pushed back on the dream of natural-language mixed-initiative conversational machines, and engaged in the most realistic

[4] http://www.darpa.mil/ipto/programs/ears/

[5] SpeechWorks was acquired by Scansoft in 2003.

proposition of directed-dialog with a meticulously engineered user interface. SpeechWorks and Nuance's goal was to build *usable* and not necessarily *anthropomorphic* systems.

The first telephony conversational applications showed business value, and attracted the attention of other players in the industry. Travel, Telecom, and Financial industries were the early adapters of directed dialog systems and deployed widely by SpeechWorks and Nuance, while the *holy-grail* of natural language communication remained in the research community. UPS, FedEx, American and United Airlines, E-trade, and Schwab were amongst the early adapters to speech enable their non-revenue generating lines of business. In the late 1990s, analysts predicted that speech market (including hardware, software, and services) will soon become a multi billion dollar business.

The concept of properly engineered directed-dialog speech applications became an effective replacement for touch-tone IVRs, and an enabler for customer self-service. A new professional figure emerged, the Voice User Interface (VUI) designer. The VUI designer is responsible for the complete specification of the system behavior, the exact wording of the prompts, and what is called the *call flow,* a finite state description of the dialog. The application development methodology was then structured according to classical software engineering principles. Requirement gathering, specification, design and coding, followed by usability tests, post-deployment tuning, and analysis, enabled speech solution providers to develop scalable, high quality, commercial grade solutions with strong Return on Investment (ROI).

In the early 2000s, the pull of the market towards commercial deployment of more sophisticated systems, and the push of new technology developed at large research centers, like IBM and AT&T Labs, prompted the industry to move cautiously from the directed-dialog paradigm towards more sophisticated interactions. IBM successfully deployed the first commercial mixed-initiative solution with T. Rowe Price, a major mutual funds company, using natural language understanding and dialog management technologies developed under the DARPA Communicator program [15]. This type of solutions is capable of handling natural language queries, such as *I would like to transfer all of my money from ABC fund to XYZ fund* as well resolving elliptical references, such as *Make it fifty percent* and allow users to change the focus of dialog at any point in the interaction.

The technology developed at AT&T known as HMIHY (How May I Help You) [16], or *call-routing*, aims at the classification of free-form natural language utterances into a number of predefined classes. Still far from providing sophisticated language understanding capabilities, HMIHY systems proved to be extremely useful and effective for routing of incoming calls to differently skilled agents, and are becoming the standard front-end for sophisticated conversational systems.

While the technology of speech recognition was assuming a more mature structure, so was the market for its commercial exploitation. Companies like SpeechWorks and Nuance initially assumed most of the industry roles, such as technology vendors, platform integrators, and application builders. At the same time, larger companies with a long history of research in speech recognition technologies, such as AT&T and IBM, entered the market. Other smaller companies appeared with more specific roles, such as tool providers, application hosting and professional services, and the whole speech recognition market started to exhibit a clear layered structure.

As the number of companies involved in the conversational market increased, the number of deployed systems rose to hundreds per year, and the need for industrial standards quickly emerged. By ensuring interoperability of components and vendors, standards are extremely important for the industry and for growing the market. VoiceXML, a markup language for the implementation of dialog systems in browser-client architectures, based on the same http transport protocol of the visual Web, was invented in the late 1990s and became a W3C recommendation (VoiceXML 1.0[6]) in 2000, followed by VoiceXML 2.0[7] in 2004. Other standards followed, such as MRCP[8] (Media Resource Control Protocol), a protocol for the low level control of conversational resources like speech recognition and speech synthesis engines, SRGS[9] (Speech Recognition Grammar Specification), a language for the specification of context-free grammars with semantic attachments, CCXML[10] (Call Control Markup Language), a language for the control of the computer-telephony layer, and EMMA[11] (Extensible Multi Modal Annotation), a language for the representation of semantic input in speech and multi-modal systems.

5 Business Cases and Business Models of Conversational Systems

Despite of its mature appearance, the conversational solutions market is still in its infancy. Transactional conversational solutions have not yet reached mainstream: only about 5% of the IVR ports in the US are speech enabled today. Thus, the conversational speech technology market is potentially very large, but penetration is still slow. When the technology will reach a reasonable level of market penetration, possibly during the next few years, conversational access will change the dynamics of e-commerce. Telecom, banking, insurance, travel, transportation, utilities, retail, and government industries are the major potential adopters of conversational technologies. Products and services offered to consumers have become more and more complex during the past few decades, requiring the above industries and businesses to develop sophisticated support infrastructure. While interaction with a consumer could lead to increased revenue opportunity, the cost of providing support for simple inquiries and transactions would require higher investments for infrastructure and agent wages. In fact, in a typical call center, labor contributes to over 70% of the total operational cost. Enabling customer *self-service* through conversational interfaces has considerably awakened the interest of the enterprises as a means to reduce operational expenses. However, as excessive automation may actually reduce customer satisfaction, finding a good trade-off between reducing cost and maintaining the quality of customer care is extremely important. That requires extensive knowledge of the business, understanding of customer needs, as well as the implementation of best practices in the call center transformation and voice user interface design.

[6] http://www.w3.org/TR/voicexml/

[7] http://www.w3.org/TR/voicexml20/

[8] http://www.ietf.org/internet-drafts/draft-shanmugham-mrcp-06.txt

[9] http://www.w3.org/TR/speech-grammar/

[10] http://www.w3.org/TR/ccxml/

[11] http://www.w3.org/TR/emma/

The overall market is maturing slowly, primarily due to a number of false starts when the technology was not ready, or with poorly engineered highly ambitious systems. That created the perception in the marketplace that speech recognition technology is not ready, it is costly to deploy and maintain, and it's difficult to integrate with the rest of the IT infrastructure. As the industry progresses with standards and more robust technology, these negative perceptions are becoming less valid. Successful industry-specific engagements and customer education can help increase the speed of the technology acceptance curve in the marketplace. Conversational self-service, unlike the Web, is still an emerging interface and as such every customer application is special and should be analyzed and developed carefully and methodically following specific best practices.

The value proposition offered by conversational applications is mainly the return on investment (ROI) created by the reduced costs of services obtained through full or partial automation. Historically, one of the first examples of a large ROI obtained by speech technology is the deployment, by AT&T, of a simple routing system which allows choosing among different types of calls by saying *collect*, *third party billing*, *person to person*, *calling card*, or *operator* [17]. The system, which was deployed in 1992, automated more than a billion calls per year that were previously handled by operators, and is said to have saved AT&T in excess of $600 million a year.

Attainment of ROI is straightforward and easily predictable in those situations when speech recognition is introduced in call centers without automation or very limited self-service. The ROI is not always obvious when the conversational system replaces an existing touch tone IVR. However, even in those situations, a higher automation rate can be obtained by using speech technology and by a complete redesign of the user interface. Furthermore, a well designed voice interface leads to higher customer satisfaction and retention, which alone can justify the choice.

As far as the business model for speech technology is concerned, licensing of core technology is certainly the one with the highest profit margin. After the initial R&D investment, vendors would benefit from a steep revenue/cost function, since the number of sold licenses is loosely dependent on the revenue production costs, such as marketing, sales, and product improvement. As the market matures, software core technology may be commoditized in the presence of market competition. If performance of products from different vendors is comparable, differentiation will come in the form of specific content (e.g. grammars, language models, dialog modules, etc.), tools to support design, development, and tuning of applications, and integration of the software with third party IT infrastructure.

Application builders have traditionally adopted the *time-and-material* model, customers pay hourly rates for initial development of the solution and subsequent maintenance and upgrades Unfortunately, the cost and the amount of specialized resources needed for the development of conversational systems is fairly high, and this model does not scale well with the increased market demand for conversational solutions. This situation has prompted the industry to develop the concept of *pre-packaged applications*, which are offered today by a large number of speech technology vendors. Pre-packaged applications address specific vertical sectors of the industry, such as finance, banking, and health. They are configurable and customizable, and since they are typically built and tuned on prior customer engagements, deployment risks are generally lower.

We're now seeing an emerging trend towards a hosting model for conversational solutions. The trend is for small and medium size businesses as well as large enterprises. Hosting is also referred to as an on-demand/utility model where clients lease solutions, and there are many different business models for every customer budget.

Finally, the overall optimization of contact centers is another interesting model for conversational solutions. This is not a hosting, but rather an outsourcing model, where companies such as IBM, Accenture, and EDS, who specialize in running large IT organizations, manage call centers for large clients. These outsourcing deals typically achieve cost reduction through call center consolidation, and multi-channel self-service, including Web, IVR, e-mail, and chat. Conversational solutions are viewed as complementary to the Web in terms of self-service, and can often outweigh the benefits realized through the Web itself.

6 Conclusions

Automatic speech recognition research, which started more than 50 years ago, found commercial deployment within a structured and maturing market only during the past few years. The vision of building machines we can talk to as we talk to humans was not abandoned, but pushed back in favor of a more pragmatic use of the technology. What enabled the change from technology to user centered design was the realization that users do not necessarily need a full replication of human–like speech and language skills; good user experience is instrumental to market adoption. Highly engineered solutions, focused on the delivery of effective transactions, and compatible with the performance of the current speech recognition technology proved to be key to the industry of conversational technology.

References

[1] Davis, K., Biddulph, R., Balashek, S. (1952), "Automatic recognition of spoken digits", J. Acoust. Soc. Amer, V. 24, pp. 637-642.

[2] Pierce, J. R. (1969) Whither Speech Recognition, J. Acoust. Soc. Amer, V. 46, pp. 1049-1050.

[3] Klatt, D. H (1977), "Review of the ARPA Speech Understanding Project," J. Acoust. Soc. Amer., V. 62, No. 4, pp. 1345-1366.

[4] Lowerre, B., Reddy, R., (1979) 'The Harpy Speech Understanding System', in Trends in Speech Recognition, ed. W. Lea, Prentice Hall.

[5] Bellman, R. (1957), "Dynamic Programming," Princeton University Press.

[6] Vintsyuk, T.K., (1968), "Speech discrimination by dynamic programming," Kibernetika 4, 81-88.

[7] Sakoe H., Chiba S., (1971) "A dynamic-programming approach to continuous speech recognition," paper 20 c 13.Proceedings of the International Congress on Acoustics, Budapest.

[8] Jelinek F., (1976) "Continuous Speech Recognition by Statisical Methods," IEEE Proceedings V. 64, N.4, pp. 532-556.

[9] Baum L. E., (1972) "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," Inequalities, V. 3, pp. 1-8.

[10] Levinson, S. E., Rabiner, L. R. and Sondhi, M. M., (1983) "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," Bell Syst. Tech. Jour., V. 62, N. 4, Apr 1983, pp. 1035-1074.

[11] Price P., Fisher W. M., Bernstein J., Pallett D. S., (1988) "The DARPA 1,000-Word Resource Management Database for Continuous Speech Recognition," Proceedings ICASSP 1988, p. 651-654.

[12] Hirschmann, L. (1992), "Multi-site data collection for a spoken language corpus," In Proc. of the 5th DARPA Speech and Natural Language Workshop, Defense Advanced Research Projects Agency, Morgan Kaufmann.

[13] Pieraccini, R., Levin, E., (1993) "A learning approach to natural language understanding," in NATO-ASI, New Advances & Trends in Speech Recognition and Coding, Springer-Verlag, Bubion (Granada), Spain.

[14] Walker M. et al., (2002) "DARPA Communicator: Cross System Results for the 2001 evaluation," in Proc. if ICSLP 2002, V. 1, pp 269-272

[15] Papineni K., Roukos S., Ward T., (1999) "Free-Flow Dialog Management Using Forms," Proc. Eurospeech, pp. 1411-1414.

[16] Gorin A. L., Riccardi G., Wright J. H., (1997) "How May I Help You?" Speech Communication, V. 23, pp. 113-127.

[17] Cox R. V., Kamm C. A., Rabiner L. R., Shroeter J., Wilpon J. G., (2000) "Speech and Language Processing for Next-Millennium Communications Services," Proc. of the IEEE, V. 88, N. 8, Aug. 2000.