From Pipedreams to Products, and Promise!

Janet M. Baker
Saras Institute / Dibner Institute
MIT – Bldg E56-100
38 Memorial Drive
Cambridge, MA 02139 USA
HistSpch@mit.edu

Patri J. Pugliese
Saras Institute / Dibner Institute
MIT – Bldg E56-100
38 Memorial Drive
Cambridge, MA 02139 USA
HistSpch@mit.edu

Abstract

This demonstration provides a historical perspective of a number of research and commercial systems in Spoken Language Technology over the past 20+ years. A series of chronologically ordered video clips from many sources will be presented to illustrate the many steps and the tremendous progress that has been achieved over the years. The clips themselves are drawn from diverse academic and commercial research labs, product presentations, and user applications. All show systems being demonstrated or in actual use. Over 20 different laboratory systems, products, and companies are represented in this collection of video materials. Each of the clips has previously been shown publicly. The present selection primarily focuses on speech and natural language systems for speech recognition and synthesis. Additional contributions to this collection are welcome.

1. Project Description

Preparations of these materials are being done in conjunction with the History of Speech and Language Technology Project being conducted by Saras Institute, in affiliation with the Dibner Institute for the History of Science and Technology, at MIT (Cambridge, MA). The overall mission for this project is to collect, preserve, and make readily available information about significant research discoveries, technical achievements, and business developments in speech and language technology. For further information on this project, please go to www.SarasInstitute.org.

Work on this project is on-going. Additional contributions of relevant materials are welcome in the area of Spoken Language Technology, including speech and natural language systems and applications incorporating speech recognition, speech synthesis, interactive dialogue, information retrieval, machine translation, multimodal interfaces, etc. Please contact us at HistSpch@mit.edu if you have materials you would like to contribute or with any inquiries, updates, corrections, and suggestions.

2. Introduction

This demonstration is intended to increase awareness and understanding of the Spoken Language Technology field, and an appreciation of the evolutionary and revolutionary steps which have turned pipedreams of the past into present products, and future promise. The progression of many stages from research and commercial laboratories into working systems are well illustrated by this collection of video clips. Each video clip (typically 1-5 minutes duration) represents technology at the cutting edge. There are 3 categories of video clips:

1) *Laboratory and Prototype Systems*
In some cases, pioneers and major contributors of the field are personally demonstrating their systems.

2) *Commercial Product Demonstrations*
These run the gamut of televised interviews/demonstrations (including many live demos) to instructional and commercial videos.

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 257–260,
New York City, June 2006.

3) *User Applications*
Here users are demonstrating how they use their systems on a routine basis in a variety of diverse applications.

3. Hardware

With the advent of ever more powerful inexpensive silicon and the integration of computers with audio interfaces, the computing platforms on which spoken language technology resides have undergone dramatic changes, progressively reaching many more users, ever more conveniently. Over the past 30+ years, we have witnessed the transition from monolithic room-filling computers to personal, even hand-held devices supporting state-of-the-art speech technology. The type and scope of applications have concomitantly multiplied with the ready access of affordable useful technology. Like other initially expensive centralized hardware and even biological systems (!), as the cost curves come down, evolution progresses, and more processing can be conducted relatively or wholly autonomously, the processing itself can be far more distributed. So while there continue to be operations or services (e.g. call centers) which are still best done in a centralized fashion, the distribution and proliferation of stand-alone systems (e.g. PCs, and cell phones) become progressively more feasible and popular.

4. Speech Synthesis

The primary focus of the present set of demonstrations is on speech recognition and synthesis. Starting with Homer Dudley's Voder (a manually-controlled speech synthesizer) at the 1939 World's Fair, audio examples of historical speech synthesis approaches and techniques clearly demonstrate extensive progress (see Resources). Video clips illustrate users interacting with different types of synthesis systems and applications.

Mechanical speaking machines in the 1700's eventually gave way to electrical devices in the early 1920's, which in turn gave way to computer generation of speech by the 1960's. The invention of the speech spectrogram in the 1940's spurred in-depth speech research, and significantly facilitated speech signal and waveform analyses, which blossomed in the 1960's.

Speak and Spell, model 2
(Owned by J. M. Baker)

Speech generation has taken two basic forms. With analysis-resynthesis the speech waveform is first parameterized and then regenerated or played back, as in the Voder. With the advent of powerful inexpensive microprocessors and DSP chips in the 1970's, a multitude of diverse sound-producing consumer products hit the market. The popular Speak 'n Spell learning toy introduced in 1978, was the most notable early entry.

Articulatory synthesis, first demonstrated in 1958, models the components and the characteristics of the physical production system - the articulators, their movements and their trajectories, as well as the vocal tract, its resonances, excitations, etc. A clear understanding of such a system is highly desirable, and may eventually be achieved. Unfortunately, the underlying complexity of the speech production system still confounds understanding and application utility. Consequently, while studies in articulatory synthesis are still ongoing, they were largely superceded by formant synthesizers.

Formant synthesis, also referred to as synthesis-by-rule, characterizes speech in terms of a source-filter model. In this model, one or more sound sources, representing the vibrating vocal cords and noise dynamically produced at articulator constrictions, excite one or more filters, representing the vocal tract and side-tube (e.g. nasal branch, etc.) resonances. A catalog of sounds (corresponding to (sub) phones, diphones, or other units) can be constructed and then reassembled (and smoothed) in accordance with a dictionary of word or phrase pronunciations. With careful selection of materials, and the careful tuning and adjustments of parameters, synthetic speech can be made to sound very natural. Although computationally efficient, automatically achieving high quality output is neither easy nor consistently achievable.

Concatenative synthesis refers to the process of sequentially combining prerecorded exemplars of speech or other waveforms to produce the desired output. A large database (including many phonetic elements, allophones, words, etc.) of well-concatenated sound elements can easily produce synthetic speech indistinguishable from natural speech. This process is very memory-intensive, but typically produces the highest quality speech synthesis available. It is widely deployed in applications requiring natural-sounding speech output from a given voice.

Although three major approaches are outlined here, a number of hybrid synthesizers, HMM synthesizers, and other synthesis methodologies are utilized to address different requirements.

5. Speech Recognition

In the past century, speech recognition has progressed from recognizing small vocabularies to transcribing general purpose dictation in real time, recognizing commands in noisy environments, and reliably extracting words and information from telephone conversations and television broadcasts. In 1922, the toy dog "Rex", would spring from his doghouse when he was called by name! Early digit recognizers were demonstrated in the 1950's and 1960's, when the predominant approach was to recognize whole word templates. This approach continued up to the beginning of the 1970's when it started being gradually replaced by Hidden Markov Model (HMM) systems using stochastic models to more accurately characterize the naturally highly variable speech signal.

Radio Rex in his house (Photo by Hy Murveit, Rex owned by Michael Cohen)

Spurred on by government funders for "speech understanding" in Europe, Japan, and the USA, many university and commercial laboratories commenced to advance the technology. Proceeding first through limited vocabulary systems and highly constrained grammars, systems gradually expanded the number of words they could simultaneously distinguish, despite greater variability in speakers, languages, and progressively more challenging acoustic/channel and natural language environments. Starting in the 1970's, hefty special-purpose commercial hardware systems were deployed for limited vocabulary industrial applications (e.g. for hands-free command/control, data entry, quality control, etc.), speaker verification, simple telephone data input and query systems, etc. Inexpensive PC sound board enabling discrete recognition started appearing in the market by the late 1970's to mid 1980's.

In the latter 1980's, vocabulary expanded to several thousand words, and then in 1990, suddenly exploded to full general-purpose dictation capabilities, though still limited to *discrete* "one word at a time" input. Meanwhile in the early 1990's, the first speech audio-mining and audio information retrieval capabilities were successfully proven to work on prerecorded *conversational continuous* speech, on telephone and broadcast data. Real-time *continuous* speech dictation "software only" products became available and sold to millions of customers by the end of the 1990's. The availability of large corpora of recorded speech and text dramatically improved modeling capabilities and system performance for both dictation and transcription.

Meantime telephone query systems allowed for users to engage in dialogue to get stock quotes, weather updates, and even make train and plane reservations. By the year 2000, commercial telephone directory assistance systems started appearing as well. Today call centers routinely employ speech technology to elicit and supply customer information through interactive spoken dialogue, in large part replacing expensive human operators. Though directed dialogues predominate, some mixed initiative dialogues ("How can I help you?") are becoming available. Spoken language translation programs, coupling speech recognition to machine translation software, started appearing first in the laboratory in the late 1990's, and then progressed to the marketplace on PCs and handheld devices, by about 2000. Major government-funded research initiatives are presently focusing on speech-to-speech systems with different language inputs and outputs.

With embedded systems, speech input and output are now available on a growing number of consumer products including automotive navigation systems, PDAs and toys; hands-free voicedialing is shipping on millions of cell phones. Despite funding cuts and other setbacks about year 2000, the steady stream of ever improving speech technology is gradually becoming an integral part of systems and services, large and small. The issues we face in improving speech technology continue to be very challenging. The retrospective afforded by this demonstration reflects on the great progress that we have made!

6. Resources

The American Association for Artificial Intelligence (AAAI), founded in 1979, includes on their website an variety of materials on the history of speech technology. http://www.aaai.org/AITopics/html/speech.html#readon

Janet M. Baker, "Milestones in Speech Technology – Past and Future!", Speech Technology Magazine, September/October 2005.

The web site "comp.speech Frequently Asked Questions" provides a myriad of links to sites dealing with the various aspects of speech technology. http://www.speech.cs.cmu.edu/comp.speech/

Eurovamp (Voice Adapted Multipurpose Peripherals) maintains a web site which includes tutorials on the basic principals and history of speech recognition and synthesis. From their Home page, click on "Training" to get to the tutorial menu. http://www.eurovamp.com/

IEEE History Center has a website on Automatic Speech Synthesis & Recognition which includes interviews with seminal contributors (under "Archives") to speech technology an other relevant materials. http://www.ieee.org/organizations/history_center/sloan/ASSR/assr_index.html

Saras Institute has a website for the History of Speech and Language Technology with information on historical artifacts, institutions, major contributors, resources, etc. http://www.SarasInstitute.org

The Smithsonian Speech Synthesis History Project (SSSHP) provides a collection of tape recordings, technical records and artifacts of speech synthesis technology from 1922 to the mid-1980s. http://www.mindspring.com/~ssshp/ssshp_cd/ss_home.htm

Special Workshop in Maui (SWIM): Lectures by Masters in Speech Processing, January 11-14, 2004, included a series of papers by senior researchers on various aspects of the history of speech technology. A "Program Guide" with summaries of these "Peer Lectures" is available at: http://dspincars.sdsu.edu/swim/WorkshopGuide.pdf

"The Typewriter one hundred years hence": cartoon from *The Illustrated Phonographic World* (June, 1984).