EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 5: January 27, 1995

Lecturer: Prof. Nelson Morgan *Scribe: Michael Shire*

5.1 Automatic Speech Recognition (ASR) History

The history of ASR is closely connected with the history of speech and audio in general. Some major historically significant stepping stones are:

- 1920's: Radio Rex
- 1950: Sonograph (Dreyfus-Graf)
- 1951: Smith, A phoneme detector
- 1952: Audrey (Daveis, Biddulph, and Balashek) - formants, digits, elaborated in 1958 with Dudley for spectra
- 1956: Olson and Belar, spectra
- 1956: Wiren-Stubbs, distinctive features
- 1959: Denes and Fry, phoneme recognizer, bigram grammar, also used rate of change

The History of ASR is also connected to research in other non-audio areas such as biological and psychological research on hearing, progress in decision theory, and progress in pattern recognition.

- 1928: Neyman & Pearson on decision theory
- 1936: Fisher on discriminant functions
- 1940's: McCulloch and Pits model for neuron

Good books on the history of pattern recognition include *Learning Machines* by Nilsson and *Pattern Recognition* by Duda and Hart.

5.1.1 Radio Rex

Radio Rex was a toy made in the 1920's. Remarkably, it was perhaps the first speech recognizer made and came before all of the major research of the 1950's. Here is a description:

"It consisted of a celluloid dog with an iron base held within its house by an electromagnet against the force of a spring. Current energizing the magnet flowed through a metal bar which was arranged to form a bridge with 2 supporting members. This bridge was sensitive to 500cps acoustic energy which vibrated it, interrupting the current and releasing the dog. The energy around 500 cps contained in the vowel of the word Rex was sufficient to trigger the device when the dog's name was called."

5.1.2 Audrey

Audrey may have been the first word recognizer. It recognized digits. It did so by approximating the formants. Formants are a way of characterizing speech and modelling the vocal tract resonances. Audrey had some important robust ideas compared to strictly observing the input spectrum. It tracked the positions of the formants instead of the energy. This is a good idea because frequency energy of the received speech signal changes constantly. Simple turning of one's head away from a direct path to the listener produces marked changes in the spectrum of the received speech. Timing information is lost but the method is still robust. Although the idea was good, there was insufficient technology to develop it completely. It worked very well achieving 2% error.

People usually have the first formant less than 900 Hz. The second formant is generally greater than 900 Hz. This is not always true, especially for small children, but to a first approximation this helps in finding the first two formants. The system (see Davis Fig.1 pg 638, pg 88 in binding) worked generally as follows. The high and low frequencies were separated and clipped. The zero crossings were counted and the formants f1 and f2 were quantized. A grid of 30 squares connected to capacitors and weighting resistors gave information similar to a correlation. The weights in conductances indicated how long an utterance spent in a square. The digits had distinguishable trajectories and so could be discriminated from one another. (See Davis Fig.2 pg 639, pg 89 in binding)

5.2 The 1950's

In 1956 Wiren used distinctive features defined by linguistics. Decisions were based on classifications such as whether the input was voiced or unvoiced, turbulent or non-turbulent, vowel or vowel-like, and so forth.

In 1958 Dudley made a classifier that continuously evaluated spectra. It did not use formants. This new paradigm was commonly used afterwards.

In 1959 Denes from the College of London added grammar probabilities in addition to accoustic information. For example probabilities on how often an 'a' follows a 'b' would be used in the recognition.

5.3 The 1960's

In the 1960's progress was made in ASR. Martin deployed neural networks for phoneme recognition in 1964. Neural networks technology has been slow to mature and preceded Markov models by a decade. Digit recognizers became better in the '60's. Phonetic features were researched. The major breakthroughs came at the end of the decade: Linear Predictive Coding and Dynamic Time Warp.

David and Selfridge put together a history table comparing various experiments. It was published in the proceedings of IRE in May, 1962. In general, people performed spectral tracking, detected a few words and sounds, and tested on a small number of people.

In 1969 Pierce wrote a caustic letter entitled "Whither Speech Recognition?". In it he argued that scientists were wasting time because people did not do speech recognition but rather speech understanding. He basically said "what's the point?".

5.3.1 Linear Predictive Coding (LPC)

LPC is a mathematical approach which has nice relations to acoustic tubes. There are efficient computational approaches. Makhoul wrote a good tutorial. Some developments are:

- Covariance method, Atal and Schroeder, 1968
- Autocorrelation method, maximum likelihood, Itakura and Saito, 1968
- Autocorrelation method, Markel, 1971
- Tutorial explanation, Makhoul, Proceedings of IEEE, 1975

5.3.2 Dynamic Time Warp (DTW)

DTW produces optimal timing normalization with dynamic programming. It was proposed by Sakoe and Chiba at around 1970. There was also a similar proposal by Itakura at about the same time. It is thought that Vintsyuk was the first to develop the theory in 1968. Good review articles on the subject are White in Transactions ASSP April 1976 and Rabiner and Levinson in the IEEE Transactions on Communications 1981.

5.4 1971-76 ARPA Project

The first ARPA project focused on speech understanding. The main work was done at 3 sites: System Development Corporation, CMU, and BBN. Other work was done at Lincon, SRI, and Berkeley. O'Malley at Berkeley evaluated grammars. The goal was to perform 1000-word ASR using a few speakers, connected speech, and constrained grammar with less than 10% semantic error. These were difficult requirements. The funding was $15 million. Only CMU Harpy fulfilled the goals. They used LPC segments, plenty of high level knowledge, and techniques learned from another project called Dragon which was being developed by Baker.

Dennis Klatt wrote a paper giving a critical review of the project. Dr. Morgan noted a possible contradiction in the paper. Klatt claimed the perplexity, which is a measure of how many words can go to and from a word, as 33. This is comparable to recent work with perplexities of 20 to 60. So except for the fewer speakers, it would seem that there has been no progress. There may be a misunderstanding. More recent reviews of this project refer to perplexities of 4.

5.5 Achieved by 1976

By 1976, researchers were using spectral feature vectors, LPC, and phonetic features in their recognizers. They were incorporating syntax and semantic information. Initial neural network approaches, DTW, and initial HMM work (Dragon) were developed. There were plenty of systems being built. Efforts on reducing search cost were explored. Techniques from artificial intelligence were also a focus.

5.5.1 Hidden Markov Models for Speech

The HMM approach uses mathematics from Baum and others, 1966-1972. It was applied to speech by Baker in the Dragon project (1974). It was later developed by IBM (Baker, Jalinek, Bahl, Mercer) 1976-1993. Many others picked it up by the mid-1980's. It is a statistical approach. Sequences of states represent sound. It has nice formalisms. It was used mostly for isolated words in the 70's. It is the dominant technology for both isolated word and continuous speech now whereas it was not 10 years ago.

5.6 The 80's

By the 1980's many people were concerned with the lack of standards such as corpora with which to compare and share results. People from industry (e.g. Texas Instruments and Dragon Systems) grouped with NIST (National Institute of Standards and Technology) and compiled a large standard corpora. Also in the 80's, front ends were developed using auditory models and dynamics. In PR engineering, HMMs were employed for continuous speech recognition. There was a change in the kinds of features being used by the recognizers. John Bridle proposed Mel Cepstra. The second major (D)ARPA ASR project was launched. Neural networks were again being used for ASR.

5.7 Some Lessons

Progress in ASR is not linear. Von Bekesy referred to a ASR progress as a "spiral". There were only a few major conceptual points developed; but lots of engineering. Studying history is good because one can see what's been tried and what can be tried again. The Klatt paper in the Waibel-Lee collection for the first ARPA project is recommended reading.

References

[DBB52] Davis, Biddulph and Balashek *Automatic Recognition of Spoken Digits*, Journal of the Acoustical Society of America. Vol. 24 No. 6, November 1952

[WL90] Waibel and Lee, ed. **Readings in Speech Recognition**, Morgan Kaufmann Publishers, Inc. San Mateo, California. 1990.