



**SWIM: Lectures by Masters in Speech Processing**  
**January 12 - 14, 2004, Maui, Hawaii**

**Summaries**

**Engineering Multimodal Interfaces**

**James L. Flanagan**  
**Rutgers University**

Pervasive digital networking is rapidly evolving to support broadband communication. In anticipation of a wealth of transmission capacity, researchers ponder applications that will utilize the capability, economically serve society, and drive demand for broadband services. An abiding vision is to provide the client a natural environment for communication, where sophisticated technology is invisible and no specialized training is required. Geographically-separate collaborators might then enjoy facile communication that captures much of the naturalness of face-to-face exchange. The ideal implies three-dimensional realism for sight, sound and touch modalities, along with access to distributed data bases and the assistance of animated artificial agents. Conversational interaction is expected to carry the principal burden of information exchange, with visual and haptic signaling providing important complements.

While the ideal user environment has not yet been realized, individual interface technologies have advanced sufficiently that primitive systems offer capabilities that substantially surpass the limited functionality and versatility of mouse and keyboard. Central to such systems is conversational interaction using automatic speech recognition and text-to-speech synthesis, along with hands-free sound capture and projection. Capabilities for automatic tracking of eye and gaze, face recognition, and detection of visual gesture exploit the sight modality. Concomitantly, force-feedback tactile appliqué enable grasp interaction, manual gesture and pointing. Typically, in natural communication these sensory channels are employed simultaneously and in combination. Fusion of the separate signals to estimate and interpret user intent is a major focus.

This report examines selected research in multimodal interfaces, and puts forward frontier issues critical to current progress. This assessment is made against the backdrop of experimental systems that include the Bell Labs HuMaNet effort and components of the information science projects of the National Science Foundation. Recent work with multimodal hand-held personal digital assistants is included. Finally, the report advocates research that will create a quantitative framework for “multimodal language”, similar to that which we have evolved for spoken language.

## **Maximum Likelihood Spectral Estimation and its Offsprings**

**Fumitada Itakura**  
**Nagoya University**

In Japan, Chiba & Kajiyama (1941) gave a complete quantitative account of the air space for each of the 5 Japanese vowel using paratography and X-ray. They re-created these spaces and successfully synthesized the vowels. Subsequently, electrical engineers such as Fant G. (1960) and others developed an acoustic source-filter theory of speech production. These pre-computer age works laid a firm engineering basis for speech signal processing of today. The emergence of computer availability in mid 1960's enabled us to make full use of a modern statistical concept in speech analysis, synthesis, coding and recognition.

This induced quite a few offspring, such as MLESS (the maximum likelihood estimation of speech spectrum), PARCOR (the partial auto-correlation analysis/synthesis, LSP (the line spectrum pair), and CSM (the composite sinusoidal modeling), they have been fully explored and developed to all aspects of speech processing and have formed the indispensable tools for speech processing of today.

In this talk, I will introduce the early development of statistical formulations of speech signals, based on the statistical theory of stochastic processes, which had been developed by Kolmogorov, Wiener, Cramer, Wold, Grenander, Whittle, Robinson, Burg, and Akaike, etc, during 1930's through 1960's. I will also mention the works of **Szegö**, Levinson, Geronimus, etc, whose pure mathematical results played very significant role in the development of these methods.

## **California Coding: Early Speech Processing in Santa Barbara, Manhattan Beach, and Silicon Valley 1967-1982**

**Robert M. Gray  
Stanford University**

This talk aims to sketch the historical and technical threads of the early development in California of what is now known as linear predictive coding (LPC) analysis. The focus is on the 1970s, but the story begins earlier and the narrative covers through the early 1980s. Personalities, institutions, and milestones are considered along with technical developments and interpretations.

The primary personalities considered are John Parker Burg, John D. Markel, A.H. (Steen) Gray, Jr., Danny Cohen, and Glen J. Culler. The institutions described include UCSB, SCRL, ISI, Culler Harrison Inc., and Time and Space Processing. The focal events are the first real time LPC speech communication on the ARPAnet, the first hardware LPC speech boxes, the book "Linear Prediction of Speech" by Markel and Gray, and the appearance of TI's Speak & Spell toy.

The technical threads involve several variations and interpretations of LPC and the encounters of early LPC research with the origins of the Internet and the precursors of wavelet analysis. Anecdotal stories of interactions among researchers in California and foreign (nonCalifornia) institutions such as NTT, Lincoln Lab, and TI are recounted. The talk is based on oral histories, the literature, email and conversations, and the author's memories as a peripheral participant.

## **Challenges in Speech Recognition**

**Lawrence R. Rabiner**

**Rutgers University and the University of California at Santa Barbara**

Speech recognition has matured to the point where it is now being widely applied in a range of applications including desktop dictation, cell phone name dialing, agent technology, automated operator services, telematics, call center automation and help desks.

Although the technology is often good enough for many of these applications, there remain key challenges in virtually every aspect of speech recognition that prevent the technology from being used ubiquitously in any environment, for any speaker, and for an even broader range of applications. This talk will analyze the 'Speech Circle' that enables a person to maintain a dialog with a machine using speech recognition, spoken language understanding, dialog management and spoken language generation, and finally text-to-speech synthesis, and show where significant progress has been made, and where there remain critical problems that need to be addressed and solved.

The talk will include several audio and video examples of speech recognition and speech understanding systems that have been studied in the laboratory to illustrate the challenges that remain to be solved before speech recognition is considered a solved problem.

## **The Cepstrum: Its History and Role in Speech Processing**

**Ronald W. Schafer**  
**Georgia Institute of Technology**

The term *cepstrum* was coined by Bogert, Healy and Tukey in a paper in an edited book published in 1965. The term was designed to reflect the change of perspective that results when one transforms a signal by taking the inverse Fourier transform of the logarithm of its power spectrum. Bogert et al introduced the cepstrum in the context of *detection* of echos. At about the same time, Oppenheim (Ph.D. thesis, 1964) proposed a new class of systems called homomorphic (or generalized linear) systems. A subclass of these systems (homomorphic systems for convolution) can be represented in terms of cepstrum-like operations. Schafer (Ph.D. thesis, 1968) and Oppenheim generalized the cepstrum to the *complex cepstrum*, defined as the inverse Fourier transform of the complex logarithm of the Fourier transform of the signal, and showed that the complex cepstrum could be used for *separating* (or filtering) the components of a convolution of two or more signals.

Early on, it became obvious that the cepstrum of a speech signal had many interesting and useful properties. This led many speech researchers to apply the cepstrum in areas such as pitch detection, formant estimation, speaker verification, and speech recognition.

This paper will begin with a brief introduction to the history of the cepstrum concept. Most of the talk will focus on the many ways that the cepstrum has been used in speech processing during the past 25 years. The talk will conclude with some new ideas that promise to stimulate renewed interest in the cepstrum and its applications to speech signal processing problems.

## **Mining the Airwaves**

**John Makhoul**  
**BBN Technologies-Verizon**

The fields of automatic speech recognition and language understanding by computer have gone through dramatic changes in their research paradigms during the last two decades. The data-driven, model-based approach that has gained dominance has resulted in dramatic improvements in the state of the art. As a result, applications of the technology have started to proliferate and are beginning to touch our lives in various ways. This talk presents the basic methodology that is used and touches on some of the salient applications. One of the applications that have gained a commercial foothold is that of indexing or mining audio data, with special interest in broadcast radio and television.

## **Models in Speech Communication Research**

**Hiroya Fujisaki**  
**Science University of Tokyo**

One of the important factors that contributed to the understanding of the nature of speech and the development of speech processing is the use of mathematically well-formulated models that allowed objective and quantitative representations of the observed phenomena and the underlying human mechanisms and processes, as well as computational processing of these representations for engineering purposes.

This talk will firstly review some of the important models that have been presented in the past, not necessarily restricted to speech coding but in a wider scope of speech communication and spoken language processing. Secondly, it will describe a few models from the author's own research. Thirdly, it will discuss some of the unsolved problems and the essential shortcomings that are inherent in the approaches currently prevailing. Finally, it will present a personal view on the goals and directions for future research on human-human as well as human-machine interaction mediated by spoken language and other media.

## **Remembering the Good Days at Bell Laboratories**

**Manfred R. Schroeder**  
**University of Göttingen, Germany**

The decades following World War Two are considered the best years of Bell Laboratories. Indeed, it was the time when the transistor, the solar battery and information theory were born. Other noteworthy advances include new error correcting codes, integrated circuits, and digital signal processing.

I was privileged to work at the Labs from 1954 to 1987 and I will attempt to convey, from my limited perspective, the spirit of those days at a unique place.

In addition to my personal adventures, I will touch on work in speech, computer graphics, concert hall acoustics, and underwater sound.

## **Speech Coding: From Vocoders to Code-Excited Linear Prediction**

**Bishnu S. Atal**  
**University of Washington, Seattle, WA 98195**  
**&**  
**Atal-Research**

The field of speech coding is now over 70 years old. It all started from the simple desire to transmit voice signals over telegraph cables. The availability of digital computers and the use of computers for signal processing made it possible to test complex speech coding algorithms rapidly. Two major developments, Wiener's work on prediction and Shannon's work, defining the information content of a message, established the theoretical framework for predictive coding. The introduction of linear predictive coding (LPC) started a new era in speech coding. The fundamental philosophy of speech coding went through a major shift. While earlier research in speech coding was inspired by the speech production mechanics in the human vocal tract, predictive coding based on Wiener's and Shannon's work led to a different approach to speech coding, resulting a new generation of low bit rate speech coders, such as multi-pulse and code-excited LPC. The semiconductor revolution came at the right time with faster and faster DSP chips and made linear predictive coding practical. Speech quality from traditional speech coding algorithms based on vocoders was always less than natural; the new speech coding algorithms continue to provide natural-sounding speech with higher and higher quality. Code-excited LPC has become the method of choice for most low bit rate speech coding applications.

Digital speech communication is rapidly evolving from circuit-switched to packet-switched networks to provide integrated transmission of voice, data and video signals. The new communication environment is also moving the focus of speech coding research from compression to low cost, reliable, and secure transmission of voice signals on digital networks, and provides the motivation for creating a new class of speech coding algorithms suitable for future applications.