Chapter 2. Advanced Television and Signal Processing Program

Academic and Research Staff

Professor Jae S. Lim, Dr. David Forney

Graduate Students

John G. Apostolopoulos, Shiufun Cheung, Ibrahim A. Hajjahmad, John C. Hardwick, Kyle K. Iwai, Eddie F. Lee, Peter A. Monta, Aradhana Narula, Julien J. Nicolas, Lon E. Sunshine, Chang Dong Yoo

Technical and Support Staff

Debra L. Harring, Cindy LeBlanc, Giampiero Sciutto

2.1 Introduction

The present television system was designed nearly 40 years ago. Since then, there have been significant developments in technology which are highly relevant to television industries. For example, advances in very large scale integration (VLSI) technology and signal processing theories make it economically feasible to incorporate frame-store memory and sophisticated signal processing capabilities in a television receiver. To exploit new technology in developing future television systems, Japan and Europe have established large research laboratories which are funded by the government or industry-wide consortia. Because the lack of a research laboratory in the United States was considered detrimental to the broadcasting and equipment manufacturing industries, a consortium of American companies established the Advanced Television Research Program (ATRP) at MIT in 1983.

Currently, the consortium members include ABC, Ampex, General Instrument, Kodak, Motorola, PBS, and Tektronix. The major objectives of the ATRP are:

1. To develop the theoretical and empirical basis for improving existing television systems, as well as for designing future television systems;
2. To educate MIT students through television-related research and development while motivating them toward careers in television-related industries;
3. To facilitate continuing education of scientists and engineers already working in the industry;
4. To establish a resource center where problems and proposals can be discussed and studied in detail;
5. To transfer technology developed from this program to the industries.

The research areas of the program include (1) design of a channel-compatible advanced television (ATV) system, (2) design of a receiver-compatible ATV system and digital ATV system, and (3) development of transcoding methods. Significant advances have already been made in some of these research areas. The digital ATV system we designed was tested in 1992 by the Federal Communications Commission (FCC) for its possible adoption as the U.S. HDTV standard for terrestrial broadcasting. No decision has been made yet on the HDTV standard.

In addition to research on advanced television systems, our program also includes research on speech processing. Current research topics include development of a new speech model and algorithms to enhance speech degraded by background noise.

2.2 Advanced Television Research Program

2.2.1 ATRP Facilities

The ATRP facilities are currently based on a network of eight Sun-4 and three DecStation 5000 workstations. There is approximately 14.0 GB of disk space, distributed among the various machines. Attached to one of the Sun-4s is a VTE display system with 256 MB of RAM. This display system is capable of driving the Sun-4 monitors or a 29-inch Conrac monitor at rates up to 60 frames/second. In addition to displaying high-resolution real-time sequences, the ATRP facilities include a Metheus frame buffer which drives a Sony

2K × 2K monitor. For hard copy output, the lab uses a Kodak XL7700 thermal imaging printer which can produce 2K × 2K color or black and white images on 11 inch × 11 inch photographic paper.

Other peripherals include an Exabyte 8 mm tape drive, 16-bit digital audio interface with two channels and sampling rates up to 48 kHz per channel, and "audio workstation" with power amplifier, speakers, CD player, tape deck, and other relevant equipment. Additionally, the lab has a 650 MB optical disk drive, a CD-ROM drive, and two laser printers. For preparing presentations, the ATRP facilities also include a Macintosh SE30 microcomputer, Mac IIx, and Apple LaserWriter.

We are considering installation of a fast network (FDDI) to augment the current 10 Mbps Ethernet. The new network would enable much faster data transfer to display devices, and it would more easily support large NFS transfers.

2.2.2 Video Representations for Low-bit-rate Applications

Sponsor

Advanced Television Research Program

Project Staff

John G. Apostolopoulos

Video is becoming an important aspect in many of today's applications and is expected to gain even greater importance in the near future. The large raw data rate of a video signal, together with the limited available transmission capacity in many applications, necessitates compression of the video. A number of image and video compression algorithms have been developed for high-definition television, video-conferencing, and video-phone applications. These algorithms create an efficient representation of the video, such as a motion field and error image, which is subsequently quantized and coded for transmission. Each of these algorithms has been specifically tailored for its particular application and operating parameters.

Personal communication devices (PCDs) and other future applications may have different sets of constraints. With possible power and bandwidth limitations, they could be required to encode the video at much lower bit rates. The bit rate could be variable for different applications or could depend on the particular transmission scenario (conditions) existing at the time. For example, PCDs or other portable devices may be able to adjust their transmission to optimally utilize the available channel capacity. The ability to efficiently operate at very low bit rates as well as at higher bit rates is therefore very important.

A representation of the video which provides a natural coupling among different video acquisition/display resolutions (temporal and spatial) could be very beneficial. For example, in point-to-point communications, knowledge of the characteristics of the acquisition and display devices and channel capacity could be exploited.

The attribute which is most likely of greatest importance for any video compression algorithm is a high-quality, intelligible reconstruction of the important information existing within the video. The ability to adapt the representation to exploit the instantaneous content of the video signal could improve the rendering of important information in each scene.

These issues are much more crucial for applications at low-bit-rates than those at higher bit rates. This research focuses on creating a scalable and adaptive/dynamic representation of the video signal, thereby enabling its very efficient and flexible compression.

2.2.3 Design and Implementation of a Digital Audio Compression System

Project Staff

Shiufun Cheung, Kyle K. Iwai, Peter A. Monta

Currently popular digital sound systems on the commercial market, such as the compact disc (CD) system and the digital audio tape (DAT) system, typically deliver uncompressed audio which is linearly quantized to 16 bits and sampled at a rate of either 44.1 kHz for CDs or, more universally, 48 kHz. These systems result in a raw data rate of over 1.4 Mbits per second per stereo pair. This is too high to be practical for applications such as high-definition television (HDTV) transmission and digital audio broadcast (DAB), both of which have channel bandwidth constraints. Recent developments in audio coding technology show that reduced data rate a sound quality comparable to CD digital audio could be achieved. In this project, we have designed and implemented an audio compression system which delivers perceptually transparent audio at approximately 125 kbits per second per monophonic channel. This represents a reduction factor of close to 6:1 from the uncompressed systems.

This compression scheme performs adaptive transform coding by passing the audio signal sampled at

48 kHz through a critically-sampled single-sideband filterbank, a technique also known in the literature as time domain aliasing cancellation. The dynamic bit allocation is performed on each transformed frame of samples. This allows the coder to adapt to varying characteristics in different sounds. To maintain perceptual transparency, many aspects of the algorithm rely on knowledge of the human auditory system. Key features include critical band analysis and exploitation of the perceptual interband masking model for quantization noise shaping.

The channel compatible digicipher system, a real-time implementation of four full channels of the audio coder, was completed in July 1992. As the audio subsystem of one of the system proposals, we entered this system into the competition for the United States HDTV standard. The real-time system, which employs Motorola 96002 digital signal processors, features color displays for monitoring purposes, direct digital interface to standard AES/EBU audio input and output, and optional Analog-to-Digital and Digital-to-Analog conversion. In October, the system successfully completed all the required subjective and objective tests established by the Advanced Television Testing Center.

Our efforts to improve the audio compression scheme are continuing. Under investigation are the possibility of using adaptive frame lengths for transform coding and the merits of using different signal representations such as the wavelet transform.

2.2.4 Design of an High-Definition Television Display System

Sponsor

Advanced Television Research Program

Project Staff

Eddie F. Lee

Several years ago, a video filter and display unit was built by graduate students to aid in the development of an HDTV system. This unit could read large amounts of digital video data from memory, filter the data, and display it at a high rate on a large, high-resolution monitor. Unfortunately, this display system is very complex and highly unreliable, and there is very little documentation to help diagnose any problems with the system.

This project, which was completed in May 1992, involved the design of a simpler and more reliable display system. Since little documentation was available, work was done to analyze the older system to determine its input specifications.

2.2.5 Transform Coding for High-Definition Television

Sponsor

Advanced Television Research Program

Project Staff

Ibrahim A. Hajjahmad

The field of image coding has many applications. One area is the reduction of channel bandwidth needed for image transmission systems, such as HDTV, video conferencing, and facsimile. Another area is reduction of storage requirements. One class of image coders is known as an transform image coder.[1] In transform image coding, an image is transformed to another domain more suitable for coding than the spatial domain. The transform coefficients obtained are quantized and then coded. At the receiver, the coded coefficients are decoded and then inverse transformed to obtain the reconstructed image.

The discrete cosine transform (DCT), a real transform with two important properties that make it very useful in image coding, has shown promising results.[2] In the energy compaction property a large amount of energy is concentrated in a small fraction of the transform coefficients (typically low frequency components). This property allows us to code a small fraction of the transform coefficients while sacrificing little in the way of quality and intelligibility of the coded images. In the correlation reduction property, spatial domain is a high correlation among image pixel intensities. The DCT reduces this correlation so that redundant information does not require coding.

Current research is investigating use of the DCT for bandwidth compression. New adaptive techniques are also being studied for quantization and bit allocation that can further reduce the bit rate without reducing image quality and intelligibility.

---

1 J.S. Lim, *Two-Dimensional Signal and Image Processing.* Englewood Cliffs, New Jersey: Prentice Hall, 1990; R.J. Clarke, *Transform Coding of Images,* London: Academic Press, 1985.

2 N. Ahmed, T. Natarajan, and K.R. Rao, "Discrete Cosine Transform," *IEEE Trans. Comput.* C-23: 90-93 (1974).

2.2.6 Video Source Coding for High-Definition Television

Sponsor

Advanced Television Research Program

Project Staff

Peter A. Monta

Efficient source coding is the enabling technology for high-definition television over the relatively narrow channels envisioned for the new service (e.g., terrestrial broadcast and cable). Coding rates are on the order of 0.3 bits/sample, and high quality is a requirement. This work focuses on developing new source coding techniques for video relating to representation of motion-compensated prediction errors, quantization and entropy coding, and other system issues.

Conventional coders represent video by using block transforms with small support (typically 8x8 pixels). Such independent blocks result in a simple scheme for switching a predictor from a motion-compensated block to a purely spatial block; this is necessary to prevent the coder from wasting capacity in some situations.

Subband coders of the multiresolution or wavelet type, with their more desirable localization properties, lack "blocking" artifacts and match better to motion-compensated prediction errors. Since the blocks overlap, this complicates this process of switching predictors. A novel predictive coding scheme is proposed in which subband coders can combine the benefits of good representation and flexible adaptive prediction.

Source-adaptive coding is a way for HDTV systems to support a more general imaging model than conventional television. With a source coder that can adapt to different spatial resolutions, frame rates, and coding rates, the system can then make tradeoffs among the various imagery types (for example, 60 frames/s video, 24 frames/s film, highly detailed still images, etc.). In general, this effort makes HDTV more of an image transport system rather than a least-common-denominator format to which all sources must either adhere or be hacked to fit. These techniques are also applicable to NTSC to some extent; one result is an algorithm for improved chrominance separation for the case of "3-2" NTSC, that is, NTSC upsampled from film.

Other work includes design and implementation of a high-fidelity audio source coder operating at 125 kb/s per monophonic channel. The coder uses results from psychoacoustics to minimize perceived quantization-noise loudness. This system forms part of the hardware submitted by MIT and General Instrument for the United States HDTV standards process.

2.2.7 Error Concealment for an All-Digital High-Definition Television System

Sponsor

Advanced Television Research Program

Project Staff

Aradhana Narula

Broadcasting high-definition television (HDTV) requires transmission of an enormous amount of information within a highly restricted bandwidth channel. Adhering to channel constraints necessitates use of an efficient coding scheme to compress data. However, compressing data dramatically increases the effect of channel errors. In the uncompressed video representation, a single channel error affects only one pixel in the received image. In the compressed format, a channel error affects a block of pixels in the reconstructed image, perhaps even an entire frame.

One way to combat the effect of channel errors is to add well-structured redundancy to the data through channel coding. Error correction schemes generally, however, require transmitting a significant number of additional bits. For a visual product like HDTV, it may not be necessary to correct all errors. Instead, removing the subjective effects of channel errors using error concealment techniques may be sufficient, and these techniques require fewer additional bits for implementation. Error concealment may also be used in conjunction with error correction coding. For example, it may be used to conceal errors which the error correction codes can detect but not correct.

Error concealment techniques take advantage of the inherent spatial and temporal redundancy within transmitted data to remove the subjective effects of these errors once their location has been determined. In this research, error concealment techniques were developed and analyzed to help protect the system from errors occurring in several parameters transmitted for HDTV images. Specifically, error concealment in the motion vectors and the discrete cosine transform (DCT) coefficients was investigated.

2.2.8 Transmission of High-Definition Television Signals in a Terrestrial Broadcast Environment

Sponsor

Advanced Television Research Program

Project Staff

Julien J. Nicolas

High-definition television systems currently being developed for broadcast applications require 15-20 Mbps to yield good quality images for approximately twice the horizontal and vertical resolutions of the current NTSC standard. Efficient transmission techniques must be found to deliver this signal to a maximum number of receivers while respecting limitations stipulated by the FCC for over-the-air transmission. This research focuses on the principles that should guide the design of such transmission systems.

The major constraints related to the transmission of broadcast HDTV include (1) bandwidth limitation (6 MHz, identical to NTSC); (2) requirement for simultaneous transmission of both NTSC and HDTV signals on two different channels (Simulcast approach); and (3) tight control of the interference effects between NTSC and HDTV, particularly when the signals are sharing the same frequency bands. Other considerations include complexity and cost issues of the receivers, degradation of the signal as a function of range, etc.

A number of ideas are currently being studied. Most systems proposed to date use some form of forward error-correction to combat channel noise and interference from other signals. Overhead data reserved for the error-correction schemes represents up to 30 percent of the total data, and it is therefore worthwhile trying to optimize these schemes. Our current work is focusing on the use of combined modulation/coding schemes capable of exploiting the specific features of the broadcast channel and the interference signals. Other areas of interest include use of combined source/channel coding schemes for HDTV applications and multi-resolution coded modulation schemes.

2.2.9 Fractal Image Compression

Sponsor

Advanced Television Research Program

Project Staff

Lon E. Sunshine

Image compression using transform coding has been a wide area of research over the last few years. The wavelet transform, in particular, decomposes a signal in terms of basis functions which are scaled versions of one another. This allows us to exploit the scale invariance or self-similarity within an image to achieve effective compression.

The self-similarity and pattern within images is a quality which may allow for high compression ratios with little quality degradation. Iterated function systems (IFS) have been shown to synthesize many self-similar (fractal) images with very few parameters. We are investigating the feasibility of using IFSs and their variants to exploit the self-similarity in arbitrary images in order to represent these images reliably with few parameters.

2.3 Speech Signal Processing

2.3.1 A Dual Excitation Speech Model

Sponsor

U.S. Navy - Office of Naval Research
Contract N00014-89-J-1489

Project Staff

John C. Hardwick

One class of speech analysis/synthesis systems (vocoders) which have been extensively studied and used in practice are based on an underlying model of speech. Even though traditional vocoders have been quite successful in synthesizing intelligible speech, they have not successfully synthesized high quality speech. The multi-band excitation (MBE) speech model, introduced by Griffin, improves the quality of vocoder speech through the use of a series of frequency dependent voiced/unvoiced decisions. The MBE speech model, however, still results in a loss of quality as compared to the original speech. This degradation is caused in part by the voiced/unvoiced decision process.

A large number of frequency regions contain a substantial amount of both voiced and unvoiced energy. If a region of this type is declared voiced, then a tonal or hollow quality is added to the synthesized speech. Similarly, if the region is declared unvoiced, then additional noise occurs in the synthesized speech. As the signal-to-noise ratio decreases, classification of speech as either voiced or unvoiced becomes more difficult, and, consequently, degradation is increased.

The dual excitation (DE) speech model, due to its dual excitation and filter structure, has been proposed in response to these problems. The DE speech model is a generalization of most previous speech models, and, with proper selection of model parameters, it reduces to either the MBE speech model or to a variety of more traditional speech models.

Current research is examining the use of this model for speech enhancement, time scale modification, and bandwidth compression. Additional areas of study include further refinements to the model and improvements in the estimation algorithms.

2.3.2 Speech Enhancement Using the Dual Excitation Model

Sponsor

U.S. Navy - Office of Naval Research
Contract N00014-89-J-1489

Project Staff

Chang Dong Yoo

Degradation caused by additive wideband acoustic noise is common in many communication systems where the disturbance varies from low-level office noise in a normal phone conversation to high-volume engine noise in a helicopter or an airplane. In general, addition of noise reduces intelligibility and introduces listener fatigue. Consequently, it is desirable to develop an automated speech enhancement procedure for removing this type of noise from the speech signal.

Many different types of speech enhancement systems have been proposed and tested. The performance of these systems depends upon the type of noise they are designed to remove and the information which they require about the noise. The focus of our work has been on the removal of wideband noise when only a single signal consisting of the sum of the speech and noise is available for processing.

Due to the complexity of the speech signal and the limitations inherent in many previous speech models, model-based speech analysis/synthesis systems are rarely used for speech enhancement. Typically, model-based speech enhancement systems introduce artifacts into the speech and the quality degrades as the signal-to-noise ratio decreases. As a consequence, most speech enhancement systems to date have attempted to process the speech waveform directly without relying on an underlying speech model.

One common speech enhancement method is spectral subtraction. The basic principle behind this method is to attenuate frequency components which are likely to have a low speech-to-noise ratio while leaving frequency components which are likely to have a high speech-to-noise ratio relatively unchanged. Spectral subtraction is generally considered to be effective at reducing the apparent noise power in degraded speech. However, noise reduction is achieved at the price of speech intelligibility. Moderate amounts of noise reduction can be achieved without significant loss of intelligibility, but a large amount of noise reduction can seriously degrade intelligibility. The attenuation characteristics of spectral subtraction typically lead to a de-emphasis of unvoiced speech and high frequency formants. This property is probably one of the principal reasons for loss of intelligibility. Other distortions introduced by spectral subtraction include tonal noise.

In this work, we introduce a new speech enhancement system based on the DE speech model which overcomes some of these problems. The DE system is used to separate speech into voiced and unvoiced components. Since the acoustic background noise has characteristics which are similar to unvoiced speech, the unvoiced component will be principally composed of the unvoiced speech plus the background noise. The voiced component will be principally composed of the harmonic components of the speech signal. As a consequence, speech enhancement can be achieved through subsequent processing of the unvoiced component to reduce the apparent noise level. New processing methods have been derived which take advantage of the unique properties of the individual components to reduce distortion introduced into the processed speech.