Special Article on Mobile Multimedia Signal Processing Technologies
An Overview

The market penetration of mobile multimedia services requires the development of a communications infrastructure exemplified by IMT-2000, and the development of compatible multimedia signal-processing technologies.

This article outlines a number of signal-processing technologies in chronological order and explains the relevant system technologies.

Hirotaka Nakano and Minoru Etoh

1. Introduction

Now that we are in the 21st century, we are about to see the appearance of new services based on the International Mobile Telecommunications-2000 (IMT-2000) standard. Although there are some claims that there is no need for data compression due to the recent introduction of broadband communication channels, we must remember that bandwidth is nonetheless limited. Considering the fact that the transmission of personal information is exponentially increasing, as in the case of the World Wide Web, data compression technology will continue to play a central role in the efficient transmission of video, audio and speech data.

This article provides an overview of the speech coding technology which is central to DoCoMo's basic services at present, as well as video and audio coding technologies that are likely to be adopted in IMT-2000 services. It also addresses the system technology that integrates these coding technologies to achieve communications systems over several networks.

2. Overview of Video, Audio and Speech Coding Technologies

When storing and transmitting digital signals, nothing is more important than developing a waveform source model for that purpose. Pulse Code Modulation (PCM) refers to a technology that models each digital signal statistically as an independent sample. Differential PCM (DPCM), on the other hand, assumes that there is a correlation between neighboring samples, and encodes the difference between the original value and the predicted value with reference to the correlation between the samples. Research and development through the 1970s and the 1980s focused on the quantization and prediction technologies of PCM and DPCM. In those days, the common issue among video, audio and speech coding technologies was, without offshoots, how to accurately code the input waveforms.

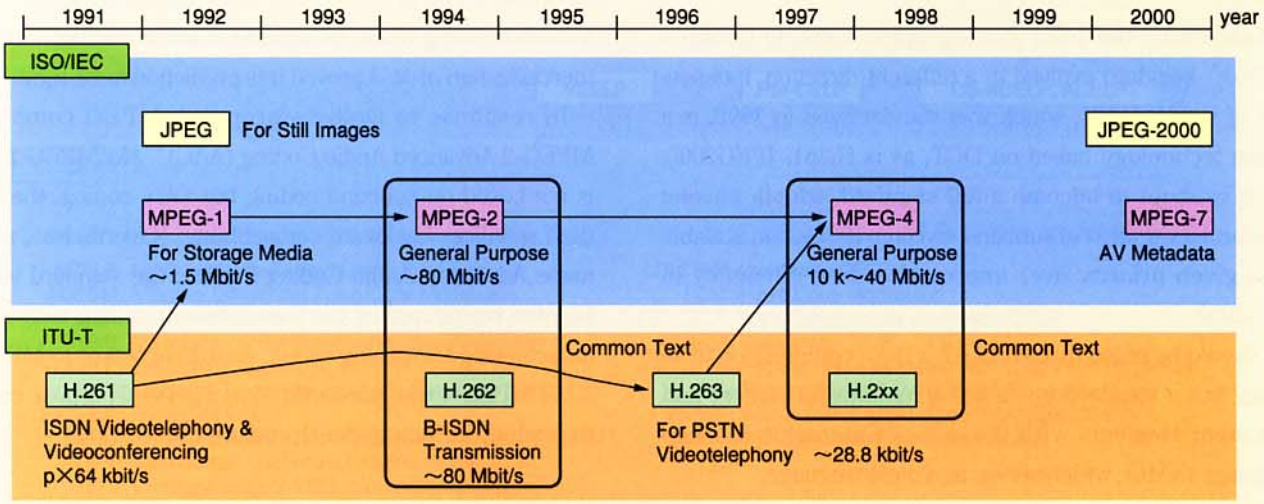
In the mid-1960s, Discrete Cosine Transform (DCT) was invented as a derivative of Discrete Fourier Transform (DFT). In the 1980s, studies on DCT's coding applications began, and subsequently, coding technologies started to diversify, along with progress in useful fundamental technologies-each being geared to a different information-source model. The following sections highlight the history of coding, in the order of video, audio and speech. But before we go any further, we must acknowledge the distinctive property of multimedia coding: the fact that media coding is a human-oriented coding technology. Video, audio and speech coding does not have to reproduce the input waveform without any loss, unless complete reconstruction is required, for example, in a medical application. In other words, distortion is permitted to the extent that humans can tolerate it. The key here-the quintessence of coding technology-is how to establish a model that defines tolerable distortions.

2.1 Video Coding

Video coding technology, as we know it today, dates back to 1990, when the standardization of the Narrowband Integrated Service Digital Network (N-ISDN) videotelephony standard H.261 was completed. H.261 is based on a combination of DCT, which takes advantage of the correlation between intra-frame pixel values, and Motion Compensation

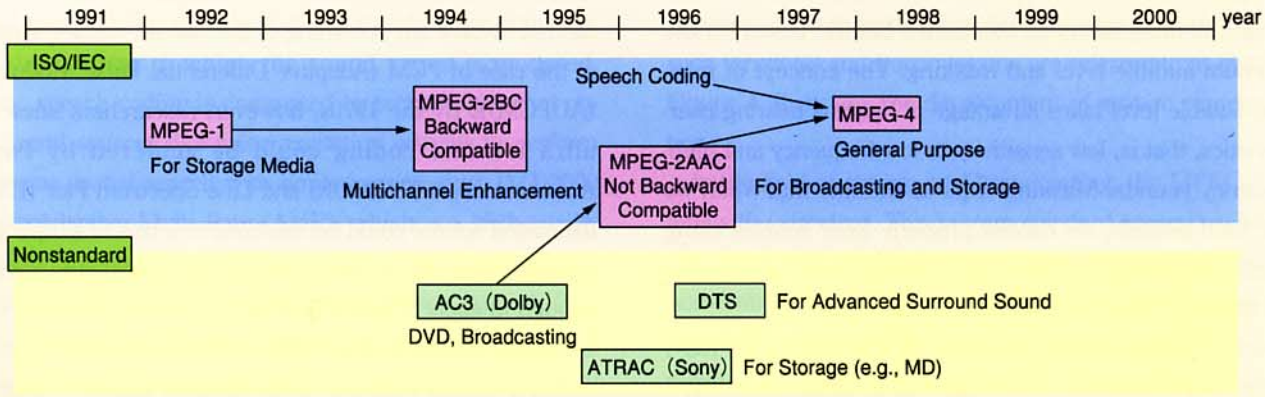
(MC), which exploits inter-frame correlations. The coding technology used in videoconferencing today is mostly H.261. In fact, the coding technologies that were developed subsequently are all offshoots of this standard, as illustrated in Figure 1. The compression ratio for video and still images is 50 : 1 and 10 : 1, respectively, depending on the redundancy of the original media.

While H.261 was designed for real-time communications, a delay is allowed in storage and broadcast applications. Exploiting this advantageous situation, the technology was enhanced to allow motion compensation based on reference to other frames. This is Moving Picture Experts Group-1 (MPEG-1), which is applied to video CDs. Then, MPEG-1 was enhanced so that interlaced video (note: analog TV systems use interlacing, in which the odd lines are transmitted on one field and the even lines on the next) could be dis-



B-ISDN: Broadband Aspects of Integrated Services Digital Network
IEC: International Electrotechnical Commission
ISDN: Integrated Services Digital Network
ISO: International Organization for Standardization
ITU-T: International Telecommunication Union-Telecommunication Standardization Sector
JPEG: Joint Photographic Experts Group
MPEG: Moving Picture Experts Group
PSTN: Public Switch Telephone Network

Figure 1 Video Coding Technologies



AAC: Advanced Audio Coding
ATRAC: Adaptive Transform Acoustic Coding
BC: Backward Compatible
DTS: Digital Theater System
IEC: International Electrotechnical Commission
ISO: International Organization for Standardization
MPEG: Moving Picture Experts Group

Figure 2 Speech Coding Technologies

played. This is MPEG-2, which enables digital television broadcasting, and many of the videos that we see today are about to be converted into the MPEG-2 format. In FY2000, DoCoMo's service network adopted MPEG-4, which is an amalgam of all coding technologies hitherto. MPEG-4 is the enhanced version of its predecessors in that it has resilience to errors arising in the wireless network, has scalability features with respect to bandwidth changes associated with Internet connection, and is highly compatible with computer graphics.

Meanwhile, the Joint Photographic Experts Group's (JPEG)★1 standard evolved in a different direction, independent of H.261. JPEG, which was standardized in 1992, is a coding technology based on DCT, as is H.261. JPEG2000, which is about to become a full standard, adopts wavelet transform-a variation of subband division. Resolution scalability is given priority over improved coding efficiency in JPEG2000.

It should be noted that MPEG-7 is not a compression technology but a standard for defining multimedia content and document elements with the use of Extensible Markup Language (XML), which serves as a meta-language.

2.2 Audio Coding

Audio coding diversified in the 1980s, as did speech coding, which is described in the following section. Audio coding is substantially different from speech coding. Speech coding evolved by specializing in the coding of speech waveforms. In contrast, audio requires the coding of uni-dimensional, continual waveforms like musical tones. Hence, audio coding involves the construction of a distortion model within human-tolerable limits as mentioned before, based on the minimum audible level and masking. The concept of minimum audible level takes advantage of humans' hearing characteristics, that is, low sensitivity to low-frequency and high-frequency sounds. Masking exploits the fact that when we hear loud sounds, we cannot properly hear sounds along adjacent frequency and time axes. These enable a compression ratio of 5 : 1 or so.

Audio coding became popular as a result of MPEG-1 Audio★2. A subsequent technique was to divide input waveforms into different subband signals through a number of filters and to allocate the optimal bit to each signal. This process is normally called subband coding. The functions of MPEG-1 Audio were enhanced following the application of MPEG-2 to video coding for broadcast purposes, i.e., it adopted multichannel and multilanguage support. The standard that focused on compatibility★3 with MPEG-1 Audio at the time is MPEG-2 BC (Backward Compatible). During the same period, Dolby Laboratories Inc. was in the course of developing Audio Code Number 3 (AC3), which did not rely on subband coding. AC3, which is based on DCT with overlapping windows, was more efficient than MPEG-2 BC, simply because subband coding is not as efficient as DCT coding. DCT is superior to subband division in terms of frequency separation, orthogonality and perfect waveform reconstruction. I had predicted that DCT coding independent of subband coding would be applied to audio coding; the commercialization of AC3 proved this prediction to be right.

In response to Dolby's proposal, MPEG completed MPEG-2 Advanced Audio Coding (AAC)★4. As MPEG-2 AAC is not based on subband coding but DCT coding, the standard sacrifices backward compatibility. Nevertheless, as the name Advanced Audio Coding implies, the standard was to become recognized as the most efficient coding technology for achieving CD-quality sound. As the successor to MPEG-2 AAC, MPEG-4 was standardized as a general-purpose coding technology including speech coding.

2.3 Speech Coding

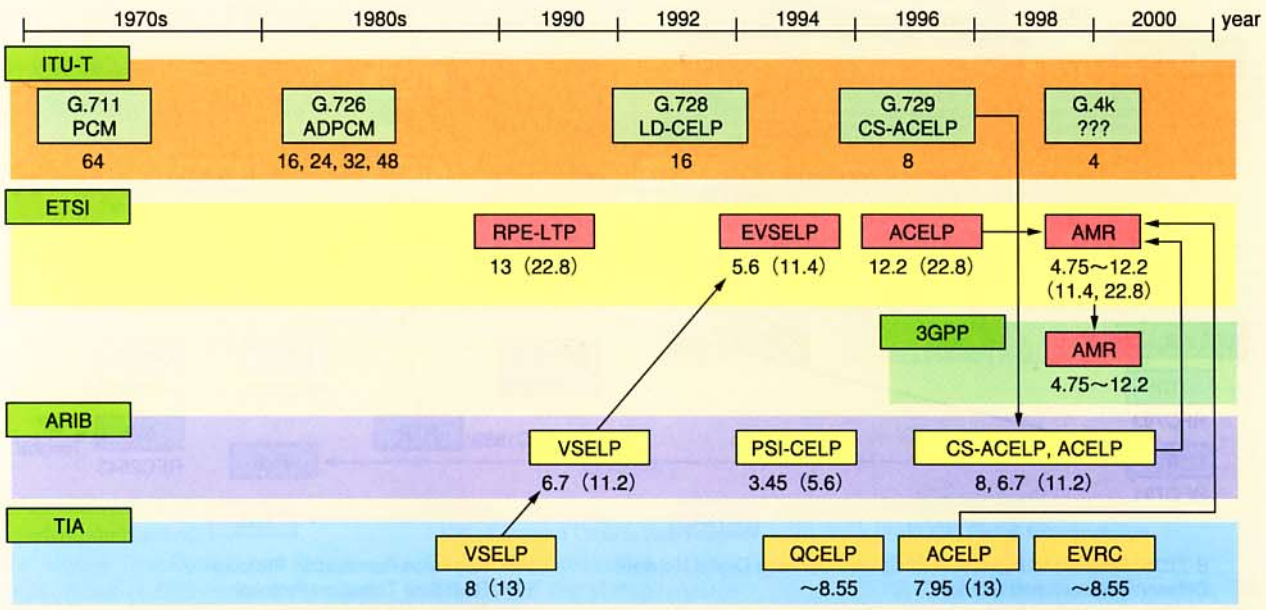
Speech coding technology specializes in coding for telephones, which is the basic communications service. In speech coding, a model of the human vocal system is built, using the vibration of the vocal cords and the modulation caused by the mouth as information source models. Figure 3 illustrates how the efficiency of speech coding has improved dramatically since the emergence of Vector Sum Excited Linear Prediction (VSELP). Before VSELP, waveform coding did not involve any modeling of the human vocal system, as in the case of PCM (Adaptive Differential Pulse Modulation (ADPCM)). By the 1970s, however, researchers knew that ultra low-rate coding could be achieved by Partial Autocorrelation (PARCOR) and Line Spectrum Pair (LSP) if the sound source could be modeled in terms of simple parameters. Then, the problem was that the quality was not high enough to apply it to practical use, as it was difficult to build a model in an environment affected by background noise and multiple sound sources. Code Excited Linear Prediction

★1 JPEG and MPEG are names of their respective organizations. However, they also became the name of the series of standards they established.

★2 MPEG is engaged in standardizing of formats to express multimedia information for storage and broadcast purposes. Hence, part of its mission is to standardize audio coding formats.

★3 Backward compatibility refers to the feature of being compatible with previous versions.

★4 MPEG-2 AAC is not compatible with AC3.



ACELP: Algebraic Code Excited Linear Prediction
ADPCM: Adaptive Differential Pulse Code Modulation
AMR: Adaptive Multi Rate
ARIB: Association of Radio Industries and Businesses
CS-ACELP: Conjugate Structure-ACELP
ETSI: European Telecommunications Standards Institute
EVSELP: Enhanced VSELP
EVRC: Enhanced Variable Rate Vocorder
ITU-T: International Telecommunication Union-Telecommunication Standardization Sector
LD-CELP: Low Delay-CELP
PCM: Pulse Code Modulation
PSI-CELP: Pitch Synchronous Innovation-CELP
QCELP: Qualcomm CELP
RPE-LTP: Regular Pulse Excitation with Long Term Prediction
TIA: Telecommunications Industry Association
VSELP: Vector Sum Excited Linear Predictive Coding
3GPP: 3rd Generation Partnership Project

Figure 3 Speech Coding Technologies

(CELP) including VSELP was developed in order to overcome the poor quality resulting from such "extreme modeling" of sound signals. CELP aimed to achieve "hybrid coding," i.e., finding a happy medium between waveform coding and the sound-source modeling mentioned above, by incorporating noise elements that cannot be modeled in the sound signal design. Most variations of CELP coding are differentiated by the way in which the sound signal is modeled. Today, speech coding is dominated by technologies that utilize sound sources like pulse excitation sources to perform extensive spatial search with limited processing. IMT-2000 adopts Adaptive Multi Rate (AMR), which is a high-quality CELP technology.

Speech coding at present has a compression ratio of 10 : 1 or so. Although its Signal to Noise Ratio (SNR) is inferior to audio coding, the perceived sound quality is not. In that sense, speech coding is efficient, based on human vocal system models.

3. System Technology

Speech, audio and video coding might form series of data called elementary streams, but that is not enough to make the communication system work. The elementary stream must adapt to the network (conversion into packets, mapping with respect to communication channels, and multiplexing for multiple elementary streams), and signaling between terminals (capability exchange and synchronization) is required. In standardization, technologies concerned with such terminals and transmissions are referred to as systems. Figure 4 illustrates the development of system standardization.

In the field of storage and broadcasting, the MPEG-2 system is the standard. This system standard prescribes how to multiplex and synchronize multiple audiovisual data for numerous digital TV programs. In the communications sector, H.320 was established by the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) in 1990, as a recommendation for videotelephony over N-ISDN. Then, H.310, H.324 and H.323 were established in 1996, subsequent to studies on terminals and systems dedicated to Broadband Aspects of Integrated Services Digital Network (B-ISDN), Public Switched Telephone Network (PSTN), and Internet Protocol (IP), respectively.

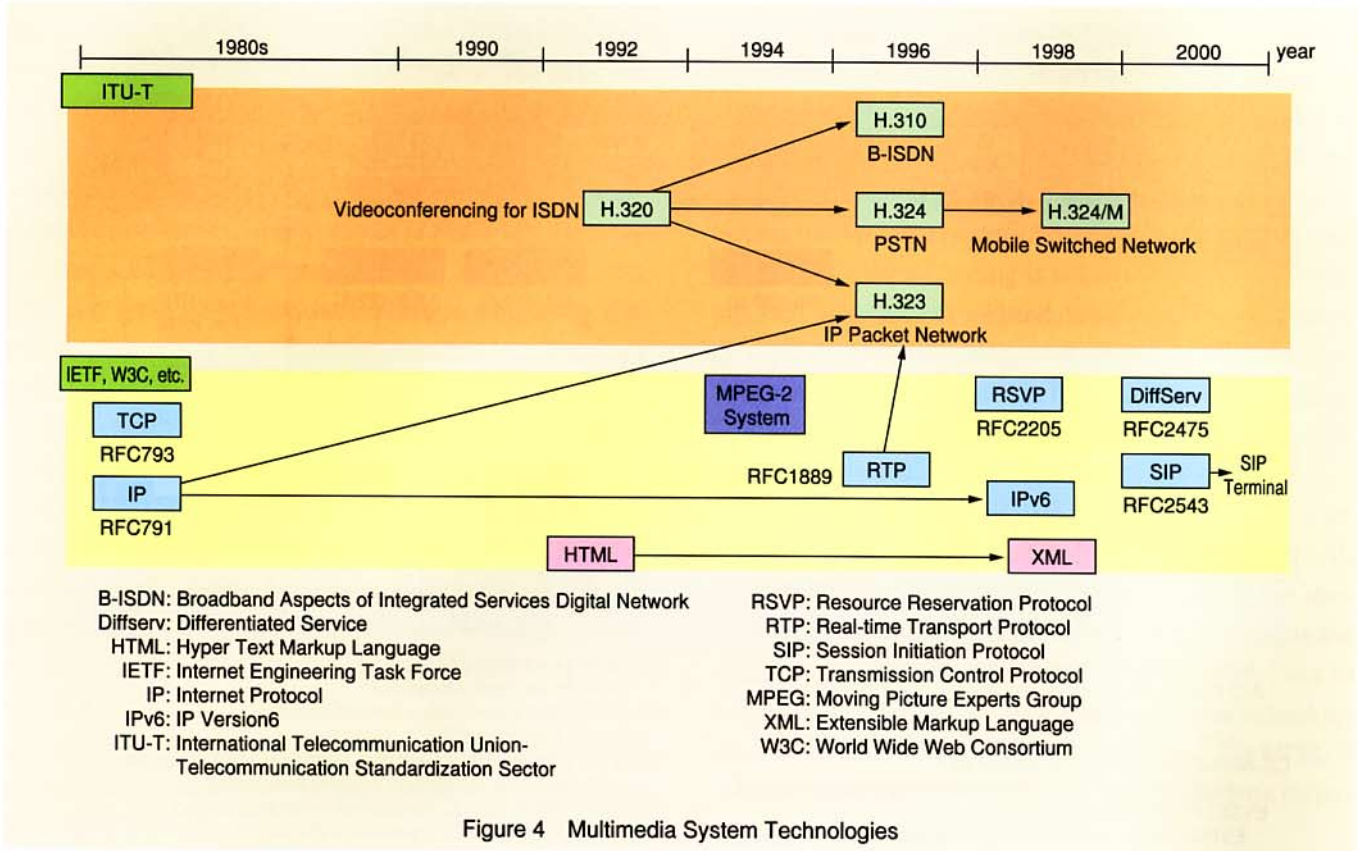


Figure 4 Multimedia System Technologies

H.324 defines the system for efficiently multiplexing and transmitting audiovisual data over PSTN. In the 3rd Generation Partnership Project (3GPP), its functions were enhanced to boost robustness against transmission errors associated with mobile communications, producing what is known as standard 3G-324M. Video terminals and visual phones that are due to become available as part of IMT-2000 services will be compliant with 3G-324M.

H.323, which defines the system specifications for IP communication terminals, is widely known as the standard for Internet phones. The Internet was originally designed as a packet switched network for data communications. Fifteen years after the establishment of the data communications protocol TCP/IP in 1981, the Real-time Transport Protocol (RTP) was established for real-time media communication. RTP aims to meet the increasing demand of multimedia data distribution in IP networks, following the recent widespread penetration of the Internet. Issues that need to be addressed to enable Internet phone services include delay and packet loss. In order to solve this, efforts are being made to apply Resource Reservation Protocol (RSVP) and Differentiated Service (DiffServ) over IP networks. Meanwhile, the Internet Engineering Task Force (IETF) is working on the Session Initiation Protocol (SIP), in order to establish a new standard for Hyper Text Transfer Protocol (HTTP) and create a system similar to H.323. As progress in system technology is dramatic, we cannot take our eyes off IETF and the World Wide Web Consortium (W3C). An example is XML, the technology that bridges systems and media. Although XML itself is a meta-markup language, which is nothing more than a language for defining a markup language, it can define the description of media (MPEG-7) and protocol, and ensure a high level of enhancement and compatibility at the same time.

4. Future Prospects

There are two issues facing multimedia coding. First, it is necessary to build a model that defines coding distortions that can be tolerated by humans and those that cannot. Although the computational complexity would be enormous, this challenge should be taken up. Second, it is essential to find an optimal way to integrate the diverse coding algorithms. A variety of algorithms exist for speech, audio and video coding, each having its own merits and demerits depending on the coding speed and information source. As a result, they are far from user-friendly. In regard to systems, efforts are being made to transfer telephone and all other media transmissions onto IP networks. Such "All-IP" efforts are also about to start in the field of wireless packet net-

works. The issue is how to handle media over IP networks, which were originally built on data communication architecture. It will be vital to discuss and decide an interface that will be in harmony with media-the payload.

5. Conclusion

This article provided an overview of mobile multimedia signal processing. A more detailed explanation is given in the subsequent special articles in this volume, which is the first in the new century.

Glossary

AAC: Advanced Audio Coding
AC3: Audio Coder Number3
ACELP: Algebraic Code Excited Linear Prediction
ADPCM: Adaptive Differential Pulse Code Modulation
AMR: Adaptive Multi Rate
ARIB: Association of Radio Industries and Businesses
ATRAC: Adaptive Transform Acoustic Coding
BC: Backward Compatible
B-ISDN: Broadband Aspects of Integrated Services Digital Network
CELP: Code Excited Linear Prediction
CS-ACELP: Conjugate Structure-ACELPDCT: Discrete Cosine Transform
DFT: Discrete Fourier Transform
Diffserv: Differentiated Service
DPCM: Differential PCM
DTS: Digital Theater System
ETSI: European Telecommunications Standards Institute
EVRC: Enhanced Variable Rate Vocorder
EVSELP: Enhanced VSELP
HTML: Hyper Text Markup Language
IEC: International Electrotechnical Commi-ssion
IETF: Internet Engineering Task Force
IMT-2000: International Mobile Telecommunications-2000
IP: Internet Protocol
IPv6: IP Version6
ISDN: Integrated Services Digital Network
ISO: International Organization for Standardization
ITU-T: International Telecommunication Union-Telecommunication Standardization Sector
JPEG: Joint Photographic Experts Group
LD-CELP: Low Delay-CELP
LSP: Line Spectrum Pair
MPEG: Moving Picture Experts Group
N-ISDN: Narrowband Integrated Service Digital Network
PARCOR: Partial Autocorrelation
PCM: Pulse Code Modulation
PSI-CELP: Pitch Synchronous Innovation-CELP
PSTN: Public Switched Telephone Network
QCELP: Qualcomm CELP
RPE-LTP: Regular Pulse Excitation with Long Term Prediction
RSVP: Resource Reservation Protocol
RTP: Real-time Transport Protocol
SIP: Session Initiation Protocol
SNR: Signal to Noise Ratio
TCP: Transmission Control Protocol
TIA: Telecommunications Industry Association
VSELP: Vector Sum Excited Linear Predictive Coding
W3C: World Wide Web Consortium
XML: Extensible Markup Language
3GPP: 3rd Generation Partnership Project