Signal Processing Technologies in Voice over IP Applications

Eli Shoval, Oren Klimker, Guy Shterlich
AudioCodes Ltd.

elish@audiocodes.com ; orenk@audiocodes.com ; guys@audiocodes.com

Abstract

In this paper, we describe some of the signal processing technologies that are required for efficient Voice over IP communication systems, the challenges that are associated with them, and how they are implemented in Audiocodes products. We describe technologies like line echo cancellation, acoustic echo cancellation and speech compression and discuss some issues relating to the implementation of these technologies in real life VoIP communication systems where both the technological and the practical aspects must be properly handled.

1. Introduction

The technology of Voice over IP is in use since the late 90's. It started as a PSTN bypass used mainly for long distance calls. It progressed steadily throughout the years capturing more applications in enterprise communication (IP PBX, UC) and carrier networks. This process happened due to the capability of this technology to be a basis for more efficient communication networks, and due to a continuous standardization effort made in recent years by standardization bodies including IETF, ITU, ETSI and others. The next step forward for this technology is going to be the adoption of wideband speech (HD) as the standard in speech communication, surpassing the capabilities of the PSTN.

In order to surpass the PSTN both in quality and in efficiency, there is a need to use sophisticated voice processing algorithms that will offer both high voice quality and efficient implementation on currently available hardware (mainly DSP processors).

In this paper, we will describe some of the basic speech processing algorithms that are required in a high quality VoIP communication system and how they are implemented in actual products – the Audiocodes AC49xx and AC50xx families of VoIP processing devices.

2. Signal Processing Technologies in VoIP

2.1 Speech Compression

The AC50xx and AC49x support a wide range of speech compression algorithms. The most basic vocoder is a G.711 in either µ--Law or A-Law format which is just a simple logarithmic PCM coder. Since the G.711 is a waveform coder, it has good performance on both speech and non-speech signals such as Fax and Modem signals. The drawback of the G.711 is its relatively high bit-rate of 64 kbps.

Low bit-rate coders can provide a good voice quality with a fraction of the G.711 bit-rate by relying on techniques such as CELP (Code Excited Linear Prediction) to model the human speech production mechanism as a linear system, where the vocal tract is represented by a set of linear prediction coefficients and its excitation is represented by a codebook. The AC50xx supports a very rich set of vocoders: G.723.1, G.729A, iLBC that are used in traditional VoIP applications, AMR, EVRC, that are used in cellular networks.

2.2 Wideband speech compression

The AC50xx and AC49x supports a set of high quality wideband speech vocoders including G.722, G.729.1 and AMR-WB [11] which enable a “better than PSTN” voice quality. These vocoders operate at the frequency range of 50-7000Hz where a common technique that is utilized in many wideband coders is to split this frequency range into 2 bands using a QMF Filter-bank and to encode each band separately. Most of the bits are allocated to the lower band, which is still the most important perceptually, while the remaining bits are allocated to the higher band. According to ITU voice quality tests of recent wideband vocoders , the quality of uncompressed wideband speech is significantly higher than narrow band speech, the difference being more than a full MOS score. The difference between a modern wideband coder such as G.729.1 at 32 kbps and a traditional Narrowband VoIP coder such as G.729A at 8 kbps is even bigger.

2.3 Line Echo Cancellation

Due to the inherent delay in VoIP systems (network delay and coding delay), these systems suffer from an echo problem where the voice signal is being echoed by the hybrid transformers that reside in CO switches and phones. The same line echo also exists in traditional PSTN calls but it is not perceived as a problem by the talker because it is being masked by their own voice. This masking is a psychoacoustic phenomenon that prevents the brain from perceiving the echo as long as the delay is small (lower than 30 ms). Since in VoIP, the delay is usually larger than 30 ms, the echo must be cancelled in order to provide a good voice quality. The AC50xx and AC49xx employs a sophisticated echo cancellation algorithm that is capable of cancelling echoes with tail lengths of up to 128 ms. This is achieved by an F I R adaptive filter that is being adapted to mimic the impulse response of the echo path (please see Fig. 1). In order to be able to support many channels per chip, the computational complexity must be constrained, the F I R filter used is a sparse filter that includes active taps only in the regions where the impulse response coefficients are substantial. The optimal locations of the active taps are being constantly estimated. The challenge in this approach is to be able to estimate these locations fast enough so the echo canceller will be able to adapt quickly to changes in the echo path. The ITU G.168 standard [3] requires that the residual echo will be reduced immediately and full adaptation will be completed within 1 sec. (please see Fig. 2), Additional challenge in echo cancellation is the performance in double talk. The echo canceller must be able to detect any double talk quickly and disable or slow the adaptation in order to prevent divergence of the filter. The AC50xx and AC49x devices use correlation metrics between the reference, input, residual and FIR output signals in order to reliably detect double talk conditions.

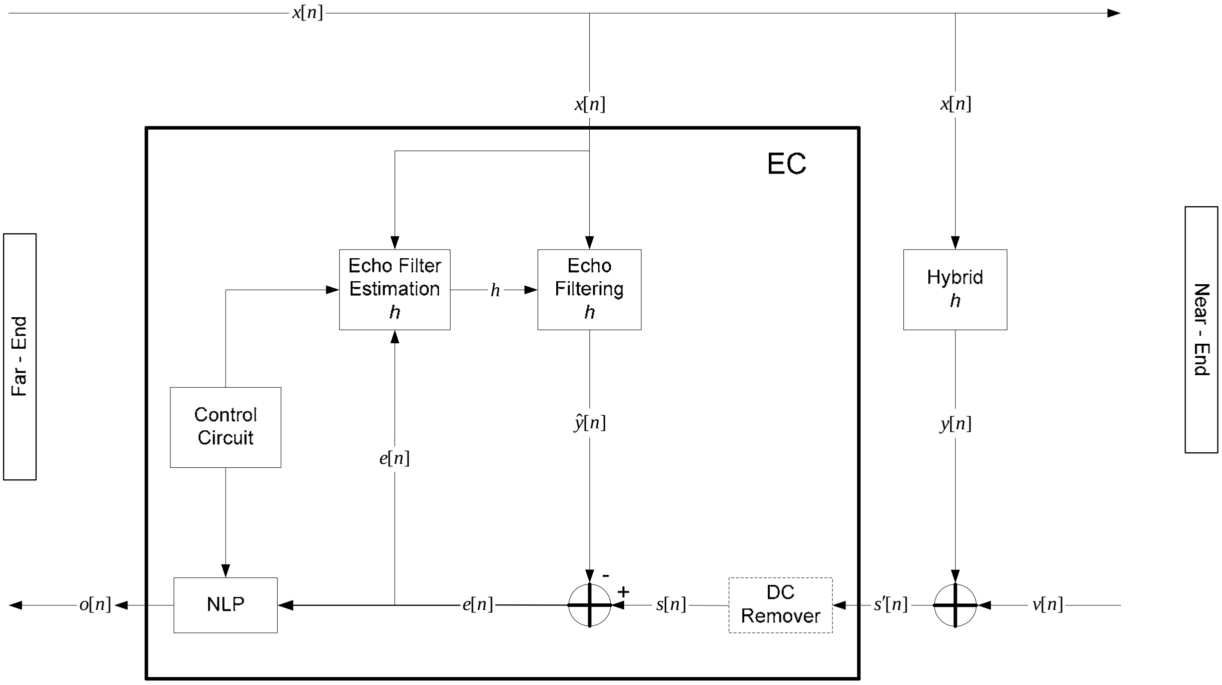

Fig. 1. Echo Canceller, simplified block diagram

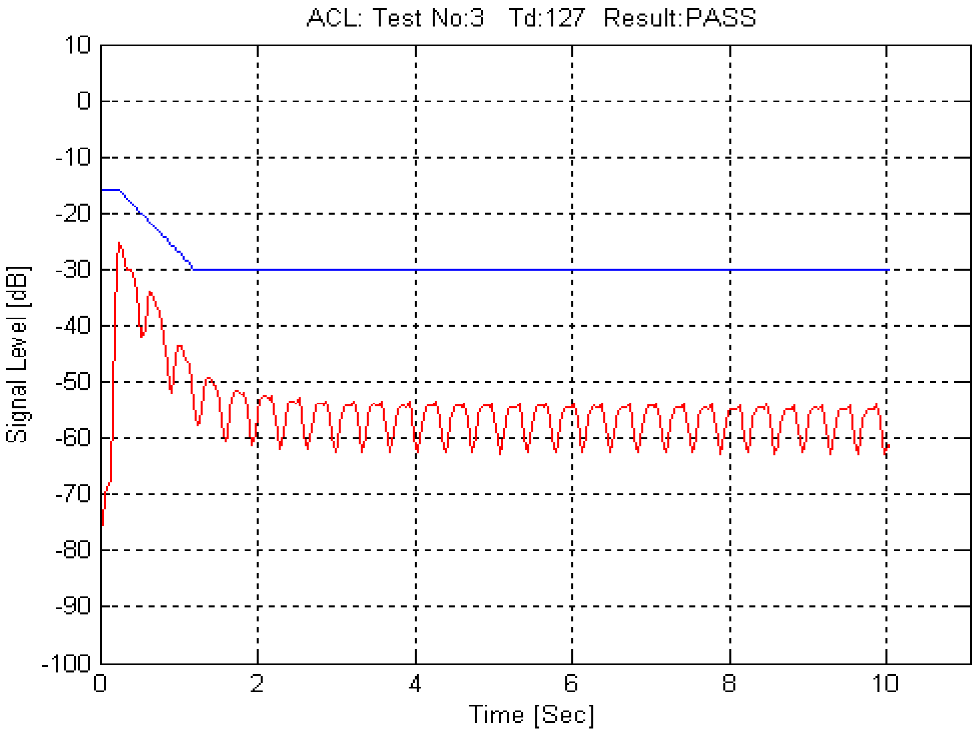

Fig. 2 Echo Canceller adaptation requirements according to test B of G.168

The upper line represents the maximal allowed residual echo as a function of time.

The bottom line represent the actual measured residual echo level.

2.4 Acoustic Echo Cancellation

The AC49x family includes the AC494 device that is used for IP Phone applications. In these applications, there is a physical problem of acoustic echo in cases where the phone operates in hands-free mode. The microphone is picking up the speech that is coming from the speaker after being reflected by walls and other objects. In order to solve this problem, the AC494 device is using an acoustic echo canceller that works in a similar way to the line echo canceller described above. However, acoustic echo cancellation presents additional challenges beyond line echo cancellation. Acoustic echo cancellation has to deal with the problems of constantly changing echo path (according to movements in the room), a worse echo return loss (i.e., the echo may be stronger than the near-end speech signal), high non-linearity in the echo path (mainly because of the speaker harmonic distortion) and a higher background noise due to the use of hands-free microphones. The AC494 acoustic echo canceller deals with the first two problems by using robust adaptation. The latter two problems are handled by a sophisticated comfort noise generator that suppresses the residual echo that is remaining due to the non-linearity in the echo path. The acoustical environment that such an IP phone is working in is characterized by a quite long impulse response of around 128 ms. This can be challenging in terms of computational burden (in wide band IP phone the FIR will have 2048 taps). So in order to handle the computational load, a sparse filter is used in a similar way as in the line echo canceller case although the number of active coefficients is larger.

3. AC50xx and AC49x families of VoIP processors General Description

3.1 Description of the VoIP processors internal hardware

Both AC49x and AC50xx families are built on silicon made by Texas Instruments. The AC491 contains 6 DSP cores of the TI C55x series, each equipped with 2 MAC (Multiply Accumulate) units running at 300 MHz. The 6 cores are each connected via an instruction cache to a shared memory of 256 K words (word size is 16-bits) as well as a local memory of 192 K words for each core. The interface to the PCM side is via 2 serial ports running at 32 MHz, each containing 512 time slots of 64 kbps. The interface to the packet side is via a Utopia level 2 interface or via a Host Port Interface. The processor is controlled via a host port of 16-bit width. Other AC49x family devices are using the same C55x DSP core with different memory structure and peripherals.

The AC50xx is using a single 64x+ family DSP core of Texas Instruments running at 700 MHz.

This core is capable of performing 4 MAC operations per cycle. It has 240KB of internal memory divided between L1 and L2 layers. It contains both HPI (Host Port) and MII (Ethernet) interface in the packets side.

3.2 Description of the Voice over IP processing flow

As can be seen in Figure 3, the voice signal enters the processor via a serial port in PCM highway format. There is also an option to process signals coming from the packet side (packet to packet feature). The signal is first filtered by a G.168 [3] echo canceller that cancels any echoes that exist in the signal due to 2/4 wire hybrid reflections. The clean signal is then classified as voice, fax, caller ID, or data modem. In the case of a voice signal, the signal can pass through an AGC and then be encoded by one of many low-bit-rate speech encoders (e.g., [9]). After encoding, the compressed bit stream is encrypted by an AES encryption algorithm [4] [5] and then packetized according to RTP standard [6]. The packets are transmitted via UTOPIA level 2 or host port or MII interface.

In the reverse direction, packets coming from the UTOPIA / host port / MII interface enter into a dynamic jitter buffer that prevents packet loss due to network jitter, while still maintaining a small delay. The packets are then decoded by the RTP decoder, decrypted, and decoded by the speech decoder. The resulting PCM signal is either transmitted via a serial port to the PCM highway or transmitted as a packet of PCM samples to the network via the UTOPIA interface.

Fax transmission is handled by the T.38 fax relay [7]. DTMF signals are handled by RFC 2833 [8] compliant DTMF relay.

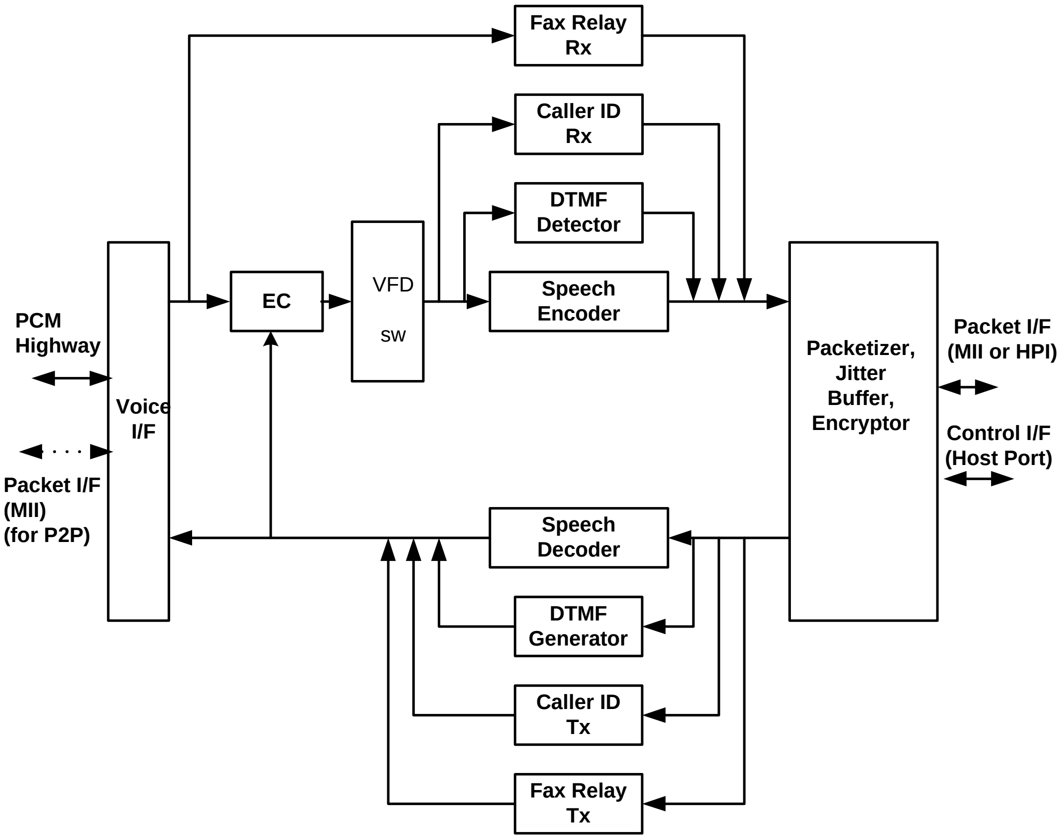

Figure 3. AC50xx Simplified Block Diagram

4. Summary

We presented some of the DSP algorithms that are used in real life VoIP communication systems and their challenges.

The AudioCodes' AC50xx and AC49x families of VoIP processors were presented. These families are scalable, ranging from 96 low bit-rate channels per device down to a single channel device. Therefore, they can cover the whole range of VoIP applications from high density media gateways and media servers to just 1 or 2 channel ATA and IP Phones. An important feature of these families is the support of wideband speech that significantly enhances the voice quality compared to the traditional narrowband speech.

5. References

[1] ETSI standard, ETS 300-019-2

[2] NEBS standard, Bellcore GR 63

[3] ITU standard G.168, Digital network echo cancellers

[4] Federal Information Processing Standards publication 197, Advanced Encryption Standard (AES)

[5] PacketCable™ Security Specification PKT-SP-SEC

[6] IETF RFC 1889, RTP: A Transport Protocol for Real-Time Applications

[7] ITU standard T.38, Procedures for real-time Group 3 facsimile communication over IP networks

[8] IETF RFC 2833, RTP Payload for DTMF Digits, Telephony Tones and Telephony Signals

[9] ITU standard G.723.1, Dual rate speech coder for multimedia communication transmitting at 5.3 and 6.3 kbps

[10] ITU standard T.30, Procedures for document facsimile transmission over the general public switched telephone Network

[11] 3GPP TS 26.171, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec

[12] ITU G.722 “7 kHz audio-coding within 64 kbps”