Audio/Telephony Integration with AC '97

revision 1.0

Written by
Dan Cox - IAL Media and Interconnect Technology Lab
dan_cox@ccm.jf.intel.com

&

Winnie Ng - DPG IHV Relations

Intel Corporation

1. Introduction

This paper is addressed to IHVs and OEMs who have detailed working knowledge of the current PC audio architecture. It is also recommended that the reader be familiar with the Audio Codec '97 Component Specification available on the Intel Web server at http://www.intel.com/pc-supp/platform/ac97/.

This paper primarily describes AC '97's support for audio functionality in relation to "Multimedia" and "Communicating" PC functionality. Implementation specific details relating to modem data pump and phone line Codec design are left to the development community, and will not be covered in this paper.

The AC '97 architecture was designed to enable vendors who choose to source both the AC '97 controller and AC '97 Codec to deliver a highly integrated audio/telephony solution. However, at this point in time, the specification of interoperability for integrated audio/telephony solutions is beyond the scope of the AC '97 definition. Vendors who wish to achieve greater levels of interoperability amongst themselves may choose to work together. The industry may wish to move toward standardization of integrated audio/telephony interoperability. Feedback to this effect can be directed to Intel at audio97@intel.com.

1.1. Headset and speakerphone connections for voice modems

The basic AC '97 supports headset and speakerphone for telephony, DSVD, and audio/video conferencing using the system audio mic and speakers. Analog speakerphone or Down Line Phone (DLP) send and receive signals can pass through the AC '97 mixer, offering the user a common point of control for all audio and telephony signals via the system audio master mixer. This also provides the user a range of mixing options, from mic only to music on hold. In the future, software support for digital mixing and interconnect may make it attractive to replace this analog interconnect.

Most voice modems employ two Codecs: one for an attached phone (DLP) and one for the Telco line. AC '97's mixer capabilities make it possible to eliminate the need for a separate phone Codec. And AC '97's optional modem DAC and ADC are designed to function as the line Codec. The optional DAC and ADC provide independent channels into the AC '97 controller, and require additional hardware support in the AC '97 controller (as well as additional software driver support).

Three telephony scenarios using AC '97 are discussed below.

1.1.1. Add-in voice modem with cabled analog connections to AC '97 system audio

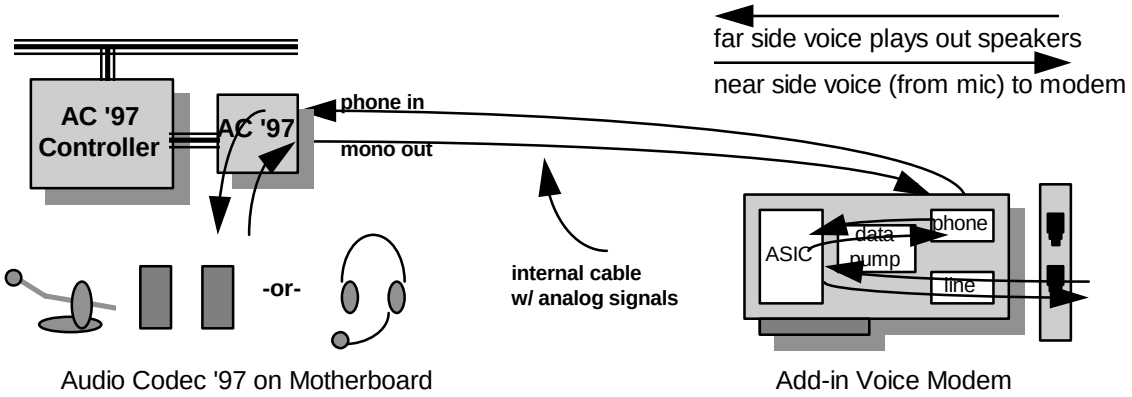

Figure 1. Voice Modem with analog connections to AC '97

For existing add-in voice modem designs with phone and line Codecs, AC ‘97’s optional modem line DAC and ADC are unnecessary. In order to utilize system audio for headset or speakerphone, the following connections are needed:

- the modem's analog **voice out** is connected to the AC ‘97 mixer's **phone in** for output through the system speakers or headset via the analog mix
- AC ‘97’s analog **mono out**, which includes the analog mic or mono mix signal (with boost and gain control), is connected to the modem's analog **voice in**

In this scenario, the modem’s phone Codec processes the voice signals, whether they come from the DLP or analog connections from AC ‘97. Echo cancellation can be performed in the modem hardware or in software on top of the phone Codec driver.

1.1.2. Add-in voice modem with digital connections to AC ‘97 system audio

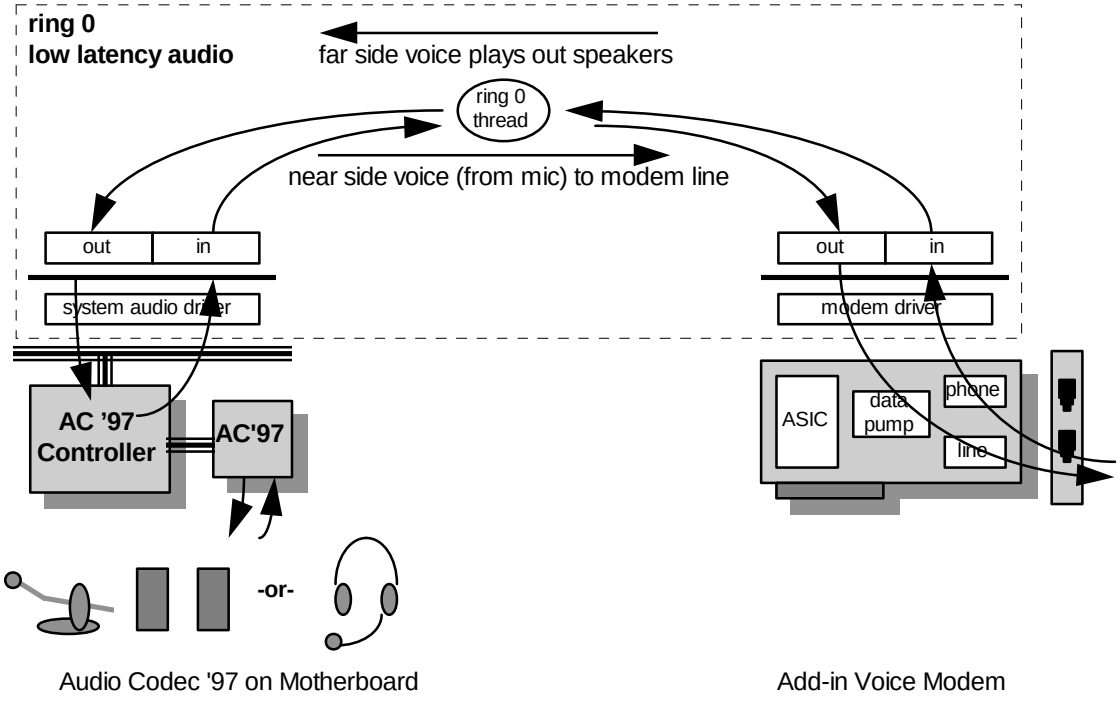

Figure 2. Voice Modem with digital connections to AC ‘97

An alternative architecture for add-in modems doesn't require any analog connections between the modem and AC ‘97. The modem still employs phone and line Codecs, but for headset or speakerphone the modem driver exchanges digital audio streams with the system audio driver via low latency ring 0 connections:

- The modem’s digital output is digitally mixed (sample rate converted if necessary) into the system audio driver’s output buffer for output through AC ‘97 to the system speakers
- The AC ‘97 mic, mono mix, or input with output (for echo cancellation) is selected for input and digitized (sample rate converted if necessary), then sent to the modem driver

In this scenario AC ‘97 functions as the phone Codec during headset or speakerphone operation, processing the voice signals. Note that the AC ‘97 mixer analog **phone in** and **mono out** signals are not used. Echo cancellation can be performed in the AC ‘97 digital controller hardware or in software on top of the system audio driver input.

1.1.3. AC ‘97 integrated audio/telephony accelerator for PCI motherboard

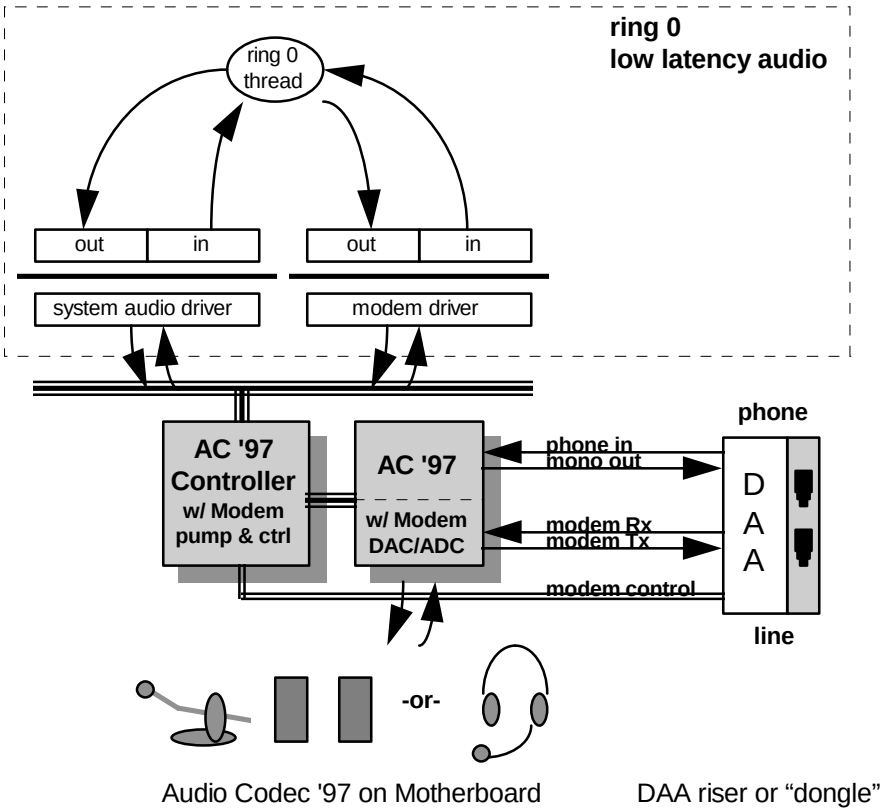

Figure 3. Integrated Voice Modem using AC ‘97

The entire telephony subsystem can be integrated onto the motherboard using the AC ‘97 design:

- Modem support (including control i/o) is integrated into the AC ‘97 Controller
- AC ‘97 system audio supports headset, or speakerphone
- AC ‘97 mixer supports play and record to/from the DLP via **phone in** (from mouth) and **mono out** (to ear) signals
- AC ‘97’s DAC and ADC function as the line Codec (**modem Rx** and **modem Tx**)

In this scenario AC ‘97 manages the DLP, mic and speakers (or headset), as well as modem line DAC and ADC. An integrated audio/telephony driver or low latency ring 0 connections are required to support full-duplex connections between system audio and the modem line Codec, which are logically independent drivers. Echo cancellation can be performed in the AC ‘97 digital controller hardware or in software on top of the system audio driver input

1.2. Speakerphone peripheral models

There are two primary peripheral models for delivering speakerphone capabilities. The selection of the model has a direct impact on the cost, capabilities and limitations of the speakerphone implementation. At this point in time, a ***dedicated speakerphone model*** is used by room conferencing equipment, and a ***system audio speakerphone model*** is used for the PC.

1.2.1 Dedicated speakerphone model

In this model the speakerphone is implemented as a dedicated mono speaker/mic peripheral. In order to apply this model to the PC, the speakerphone unit might attach to the telephony card or be connected via USB.

PROs

- allows for ideal mic/speaker placement and calibration of acoustic coupling
- the echo cancellation filter algorithm and control software can be specifically tuned to the hardware
- model supports "soft" or hardware accelerated versions of same peripheral

CONs

- Adds cost and desktop clutter due to duplication of mic and speaker resources
- speakerphone operation may be affected by other audio sources which play through the system speakers

1.2.2 System audio speakerphone model (system mic and speakers)

In this model the speakerphone is implemented with the mic and speakers attached to the PC's system audio Codec.

PROs

- saves cost and clutter, utilizes system mic and speakers
- centralized audio/telephony architecture enables cancellation of more than just voice

CONs

- the echo cancellation filter algorithm and control software must operate with a variety of desktop mics and speakers.
- speakerphone operation may be affected by other audio sources (digital or analog) which get mixed in at the Codec and play through the system speakers

1.3. Acoustic echo cancellation for speakerphone

Acoustic Echo Cancellation (AEC) filtering is a requirement for speakerphone functionality in full-duplex (simultaneous input and output) PC telephony and conferencing environments, and could become an exciting new capability for "hands free" DSVD games. AEC involves removing reflections of the output signal (which plays through the speakers) from the incoming signal (captured at the microphone). An echo cancellation filter requires digital representations of both the input and output data streams.

There are several forms of Echo Cancellation (EC) in the PC audio/telephony environment:

- ***Hybrid Echo Cancellation*** (HEC) for telephony cards with DSVD
- Half-duplex Acoustic ***Echo Suppression*** (ESP) for mono 8Kss telephony/conferencing
- Full-duplex mono ***Acoustic Echo Cancellation*** (speakerphone AEC) for 8Kss telephony/conferencing
- Full-duplex 2-channel AEC (stereo AEC) for removal of ***stereo*** output from the mic input
- Full-duplex AEC for ***mono 16Kss*** telephony and conferencing (wide-band AEC)

"Speakerphone AEC" and "stereo AEC" are discussed in the following sections.

1.3.1. Mono "speakerphone AEC"

The typical hardware or software echo canceller for "speakerphone" telephony removes audio coupling caused by the far end voice playing through the near end speakers and returning via the near end mic (heard as *echoes* on far end). This requires 1 delay line and 1 filter to remove the far end **mono out** (present in both **L out** and **R out**) component from near end **mic in**.

- ***"Speakerphone AEC"*** requires mic input + "local" mono output reference signal
- reference signal needs to be ***time correlated*** and ***same sample rate*** as mic input

"Speakerphone AEC" is primarily designed as a "voice only" echo canceller, and prevents talkers on the far end from hearing echoes of their own voice caused by acoustic coupling on the near end. This type of canceller employs an *adaptive filter* and, being modeled on the dynamics of a two way conversation, typically performs *voice activity detection* and attempts to update the filter coefficients under the right conditions. If continuous music or true stereo audio sources play, this can interfere with adaptation.

The AC '97 architecture implements hardware support for the "speakerphone AEC" function.

Depending on the actual filter implementation, there may or may not be benefit to adding L+R (or performing stereo to mono conversion) to get a better output reference signal in order to extend the cancellation capability to other than mono source material with a 1 delay line filter:

- If what is playing is actually 2-channel mono, then no harm is done to the reference signal, and the original "voice only" speakerphone echo canceller should work well.
- If stereo audio mixed with mono voice is playing, then reducing the reference signal to mono should still support removal of the mono far end voice component, and may also enable removal of a substantial amount of the stereo material as well.

NOTE: For robustness, the filter coefficient adaptation model for this type of echo canceller should be less dependent on *voice activity detection*.

1.3.2. "Stereo AEC"

Robust DSVD games with speakerphone requires a new form of echo cancellation. In a DSVD game scenario which employs speakerphone, it would be highly desirable to additionally remove the near end game's audio material (music, sound effects, explosions, etc...) playing to speakers from what is transmitted via mic to the far end (heard as *interfering material* on far end). This requires 2 delay lines and 2 filters to independently remove **L out** from **mic in** and **R out** from **mic in**. Speech recognition (for command and control) during the presence of stereo output also benefits from the "stereo AEC" filtering. The result is a mic input signal free from interfering stereo output material.

- ***"Stereo AEC"*** requires mic input + "global" stereo output reference signal
- reference signals need to be ***time correlated*** and ***same sample rate*** as mic input

NOTE: "Stereo AEC" differs from "speakerphone AEC" due to the presence of continuous music or true stereo audio sources, and requires a more sophisticated filter coefficient adaptation model.

The AC '97 architecture implements hardware support for the "stereo AEC" function.

1.3.3. Software support for echo cancellation

The current generation of host-based echo cancellers operate on a data format known as "***time correlated i/o***". This format is only meaningful when the Codec is operating in full-duplex mode. The time correlated i/o format is a 2-channel format which resembles the traditional interleaved stereo format. Each sample, instead of containing left and right inputs, contains an input (captured from mic) in the left channel and the current output (destined for speakers) in the right channel.

Audio drivers which support full-duplex Codecs should be able to correlate the Codec's **PCM out** and **PCM in** streams to produce the 2-channel time correlated i/o format. If mono PCM is playing out, the audio driver simply interleaves the output samples (which just played) with the incoming mono PCM samples (which were just collected). This technique has been successfully used to develop host-based echo cancellers running at 8Kss on a variety of Codecs.

If stereo PCM were playing out, the audio driver could add the left and right output samples (perform stereo to mono conversion) and interleave the sum with the incoming mono PCM samples. In this case, 2-channel time correlated data could be made available to the input stream for an enhanced mono "speakerphone AEC" function.

1.3.4. Hardware support for echo cancellation

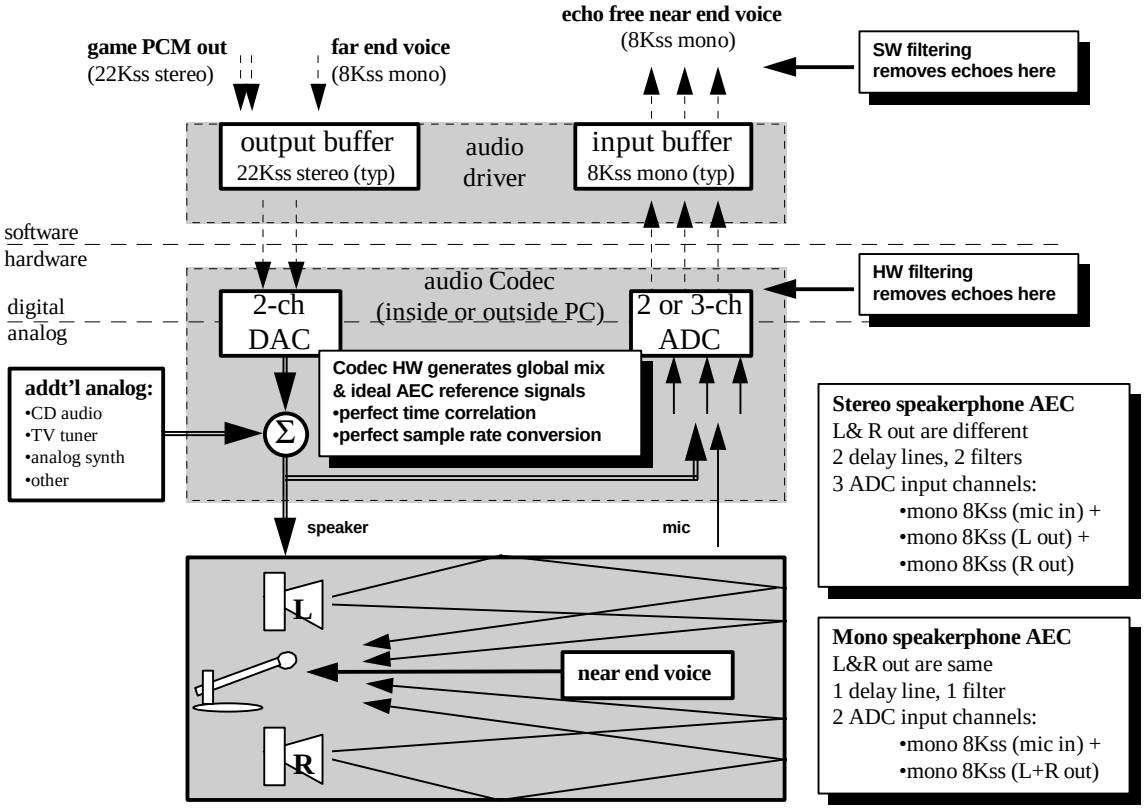

Figure 4. System diagram for Echo Cancellation

Today's multimedia Codecs generate or mix additional audio sources (digital and analog) into the output signal (such as hardware MIDI synthesis and analog Red Book CD audio). There is no way to capture this output data at the software driver level. But there are simple additions that can be made to the Codec mixer and input MUX to support the capture of 2- or 3-channel input data that is perfectly time correlated and at the desired input sample rate.

For "speakerphone AEC" the L and R output channels can be mixed together (in the analog mixer stage) to generate a mono echo cancellation reference signal. Recording 2-channel data from mic return **mic in** in the left channel and the Codec's internal generated **mono out** mix for the right channel. This technique was recommended in Intel's Audio Hardware Interface '96 Design Guide (a.k.a. Codec '96), and has been implemented in several of the currently available 1-chip ISA Multimedia Codecs. The baseline AC '97 mixer implements support for this capability.

For "stereo AEC" a 3-channel format is needed. The AC '97 analog mixer defines an optional 3rd ADC input which is dedicated to the mic, and can be used to return **mic in** along with the **stereo out**.

1.3.5. Current implementations of host based "speakerphone AEC"

Current host based full duplex (FDX) speakerphone implementations, such as Intel's controllerless DSVD modem and ProShare™ Video Conferencing System, use an echo cancellation filter which implements 1 delay line and 1 filter to remove the far end voice from what is returned via the near end mic. Since hardware support for software echo cancellation has not yet been widely implemented, Intel's "speakerphone AEC" implementation uses a native audio driver which supports software correlation. This echo cancellation filter spans the output stream which gives it control over the near end output volume (which is primarily needed in half-duplex (HDX) mode before convergence or if the canceller is told to save host MIPS by only doing HDX). Output volume control could also have been implemented via Microsoft's Mixer API.

This AEC filter supports being informed of volume control changes so as to track volume levels and minimize the need to reconverge when levels are adjusted by the user (however, users who adjust *any* speakerphone's volume control will typically notice some artifact).

Currently, the limitations of current software audio drivers' ability to generate the needed echo cancellation reference signals is inhibiting the development of more advanced echo cancellers. Developing the next generation of speakerphone and stereo echo cancellers may depend on support from the audio Codec hardware.

1.4. Modem standards

1.4.1. V.34

V.34 is the ITU-T[1] data modulation standard for modems operating at speeds up to 28.8 Kbps. A revision of V.34 which allows two new modem speed options, 31.2 Kbps and 33.6 Kbps, is expected to be ratified in October 1996. V.34 is the modem datapump standard required in DSVD 1.2 and V.70 modems, and H.324 terminals.

V.34 modems support both full duplex and half-duplex modes of operation and channel separation by echo cancellation techniques. Half-duplex mode will be used by high-speed FAX devices. V.34 modems support synchronous data signaling rates ranging from 2.4 Kbps to 28.8 Kbps (soon 33.6 Kbps) in multiples of 2.4 Kbps. The data signaling rate is determined by the modems during modem startup by measuring the transmission characteristics of the connection, and operating at the highest possible rate that will be supported by that particular connection.

1.4.2. V.70

V.70 is the ITU-T standard for DSVD modems. The standard supports the simultaneous transmission of data and digitally encoded voice signals over the POTS[2] network. V.70 supports point-to-point connections for DSVD operations. V.70 also supports multipoint connections through a Multipoint Control Unit (MCU).

A V.70 system has the following characteristics:

- the ability to enter the DSVD operating mode either at call set-up, or during an analog telephone connection
- multiplexing of bi-directional voice and data channels using a V.42 LAPM based multiplexing technique, V.76
- Transmission of the multiplexed bit stream using the modulation technique defined in V.34 or V.32*bis*

[1] International Telecommunication Union - Telecommunication Standardization Sector
[2] POTS - "Plain Old Telephone System"

1.4.3. V.80

V.80 is the ITU-T standard for the Synchronous Access Mode (SAM) DTE[3]-to-modem protocol. V.80 is a new standard protocol for the PC COM port. This standard is the key enabler of host based implementations of communication protocols on PC's. Host based communication protocols have to communicate with modem controllers and datapump through the asynchronous PC COM port. V.80 enables the bi-directional and simultaneous exchanges of commands and data between the PC host and a synchronous modem datapump. V.80 converts the synchronous bitstreams to asynchronous octets in order to pass the data through the PC COM port. This process is called Synchronous Access Mode (SAM).

V.80 allows host based software to implement new communications protocols. For example, a H.324 based video conferencing application is an example of applications that can utilize the V.80 standard. In most implementations, V.80 will not require a change to the modem hardware but rather an update to the modem controller firmware.

1.4.4. H.324

H.324 is the ITU-T standard for an analog multimedia communication terminal which utilizes V.34 modems operating over the POTS network. H.324 terminals may carry real-time voice, data, and video, or any combination thereof, simultaneously over a high speed modem connection.

H.324 terminals can be integrated into personal computers or implemented in stand-alone devices such as videophones. Support for each media type (voice, data, audio) is optional. H.324 supports simultaneous use of multiple channels of media types.

Modems used in H.324 terminals must operate in full-duplex, in either synchronous mode or V.80 synchronous access mode, and conform to ITU-T V.34 and ITU-T V.8 standards; support of V.8*bis* is also required.

1.5. Host Based Modem Implementation

1.5.1. Host based modem controller

In order to provide their customers with cost-effective communication solution such as video conferencing, many vendors have or are working on modem products with host based modem controllers (hardware data pump). AC '97 can support the audio Codec hardware requirements to provide functionality of plain data, voice and data, and video, voice and data modems.

1.5.2. Host based data pump

With the introduction of more powerful processors and Intel's MMX™ technology, some vendors are working on host based modem solution (soft modem). AC '97 can support the audio Codec requirements of a soft modem implementation by providing the i/o functionality required in modem communication environment.

[3] DTE - Data terminal equipment. It can be the host PC or modem hardware depending on implementation.