Audio Technology Used in AT&T's Terminal Equipment

John C. Baumhauer, Jr.
Scott H. Early
John H. Fikus
Steven L. Gay
Michael A. Zuniga

This paper discusses the audio technology of AT&T's terminal equipment. It describes electroacoustic and digital signal processing (DSP) techniques that improve the perceived audio quality and ease of use of telephones and speakerphones. The electroacoustic techniques include directional microphones, close-talking microphones, and loudspeakers of varying designs. The DSP techniques include echo control, speech enhancement, and active noise cancellation. These techniques increase signal-to-noise ratios, expand transmitted bandwidth, and reduce the effects of room reverberation, echo, and other common audio problems.

Introduction

The high value that AT&T's customers associate with its audio communication services is, no doubt, due to the quality of the audio they experience during their conversations. This, in turn, is dependent to a large degree on the audio quality of the terminals at each user's location—that is, their telephones and speakerphones. This paper discusses the current and evolving state of the audio technology in AT&T's terminal equipment.

Modern audio technology is based on the combination of *electroacoustics* and *digital signal processing* (DSP). Electroacoustics is the science of the interconversion of electric signals and acoustic signals using transducers, such as microphones and loudspeakers. DSP involves the manipulation of digitized signals to enhance signal quality. Our goal is to show how good audio performance can be achieved.

Applications of audio technology to telephony can generally be divided into either *handset-based* or *hands-free* systems. The example shown in Figure 1 illustrates a point-to-point conversation between two people—Frank, using a hands-free terminal, such as a speakerphone, and Mary, using a handset-based terminal, that is, a telephone. The terminals are connected over the telephone network, which is ideally modeled here as providing a delay (D) in each direction.

Frank's and Mary's perceptions of the audio quality of their conversation are influenced by two major factors: the perceived sound quality of the received voice and the ability to carry on a natural, fluent conversation. The perceived sound quality is a function of bandwidth, fidelity for the given bandwidth, and background noise from various sources. The fluency of the conversation is mainly affected by the ability of each talker to both speak and listen simultaneously, as in a *full-duplex* (see Panel 1) handset-based to handset-based connection.

Mary's perception of the sound quality of Frank's voice is adversely affected by:

- The background noise in Frank's room that is picked up by the speakerphone's microphone;
- The reverberation or *barrel effect* on Frank's voice signal due to sound reflections off of the walls of the room, and
- The attenuation of low frequency sounds at Mary's end of the connection if her handset is not pressed firmly against her ear to form a secure acoustic seal.

Frank's perception of Mary's voice is a function of:

- The effectiveness of the acoustic coupling between Mary's mouth and her handset microphone,

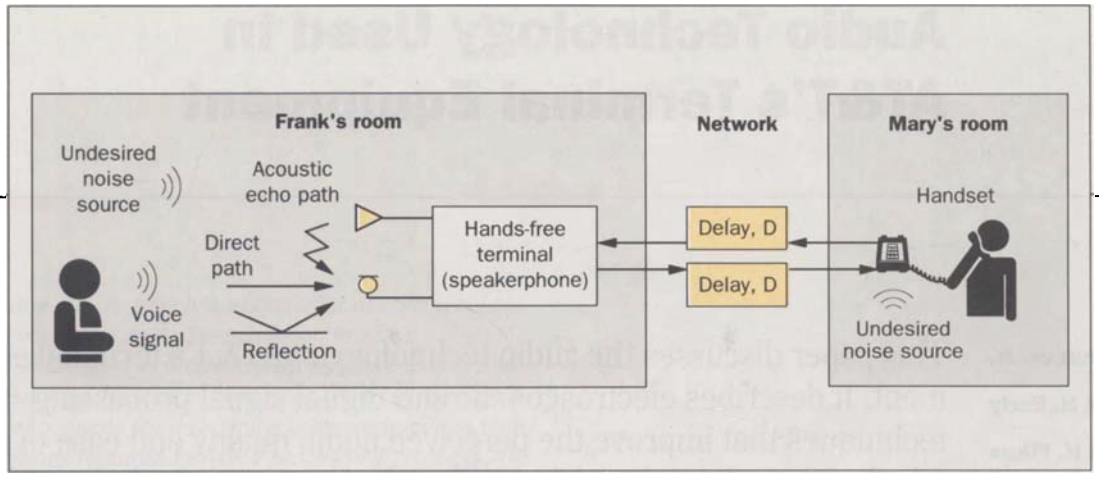

Figure 1. This illustration shows a point-to-point conversation between Frank, using a hands-free terminal, such as a speakerphone, and Mary, using a handset-based terminal, that is, a telephone. The terminals are connected over the telephone network, which provides a delay (D) in each direction.

Frank's and Mary's perceptions of the audio quality of their conversation are influenced by two major factors: the perceived sound quality of the received voice and the ability to carry on a natural, fluent conversation.

- The environmental noise picked up by Mary's handset microphone,
- The gain setting on Frank's speakerphone loudspeaker, and
- The environmental noise at Frank's location.

In addition, the fluency of Frank and Mary's conversation is adversely affected by the degree of *echo* that is passed through the system. Mary's speech, which is broadcast into Frank's room by the speakerphone loudspeaker, is nominally picked up by the speakerphone microphone and transmitted back over the network. Without some sort of echo control, Mary would hear an annoying echo of her own voice. The longer the round-trip delay ($2 \times D$) inserted by the network, the more annoying the echo.

The conventional speakerphone uses an echo control technique, called voice-activated switched-loss, which is *half duplex* in nature. This tends to prevent Frank and Mary from speaking simultaneously or interrupting the other's conversation. This unnatural pattern of conversation is disconcerting to both users.

Human/machine interactions also suffer from acoustic-related problems. For example, consider the situation where Mary in Figure 1 is replaced by a computer with a speech-synthesizer/recognizer interface. Outgoing messages from the computer to Frank will echo back and interfere with the computer's speech recognizer. A switched-loss speakerphone can eliminate the echo, but at the price of suppressing the computer's synthesized outgoing message whenever there is a sufficiently loud acoustic signal, such as a background talker, in Frank's room.

The next section of this paper, "Transducers," shows how new transducer techniques, often using DSP, can be used to solve some of the problems listed above. The third section, "Digital Signal Processors for Acoustics," shows how advanced DSP can provide improvements in the quality and naturalness of the conversation. And finally, the fourth section, "Conclusions," lists areas for further research and development.

Transducers

This section discusses microphone and loudspeaker systems designed to address the problems of handset-based and hands-free telephony. The discussion begins with some aspects of microphone technology and loudspeaker technology in handsets.

Microphone Technology. Basically, a microphone is a transducer that converts an acoustical input signal to an electrical output signal. The *directivity pattern* of a microphone is a measure of the amplitude of its electrical output, given the angle of incidence of the input, which is the acoustic wave.

Omni-directional microphones provide the same output amplitude, regardless of the direction of the input. *Directional microphones*, on the other hand, are more sensitive to acoustic waves impinging from certain directions.

Figure 2 shows the directivity patterns of three directional microphones. These are polar plots of the electrical output magnitude versus the angle of incidence for an acoustic input of constant amplitude and frequency. The plots are normalized with respect to the largest output.

Figure 2A shows *dipole* directivity, Figure 2b shows *cardioid* directivity, and Figure 2c shows *hypercardioid* directivity. For the dipole, the direction with the greatest signal input is 0° and 180°, each of which is

Panel 1. Abbreviations, Acronyms, and Terms

AEC—Acoustic echo canceler
AEP—Acoustic echo path
ANC—Active noise cancelation
AVC—Audio for video-conferencing
cm—Centimeter
dB—Decibel
DSP—Digital signal processing
DTAO—Desk-top/audio-only
Duplex—The word duplex means communication in both directions simultaneously. Full duplex is therefore redundant and half duplex is an oxymoron. However, these terms have attained common usage in the literature.
EEC—Electrical echo canceler
EEP—Electrical echo path
FIR—Finite impulse response
HMM—Hidden Markov model (See Panel 2, p. 50 of this issue)
Hz—Hertz
ML—Maximum likelihood
mm—millimeter
NLMS—Normalized least mean square
RTD—Round-trip-delay
SBEC—Subband echo canceler
SNR—Signal-to-noise ratio
SNR_e—Signal-to-electrical-noise ratio

called the microphone's *main beam*. The directivity pattern has a null at 90° and 270°, that is, the direction from which the input signal is completely rejected.

For the cardioid pattern, the direction with the greatest signal input is 0°, although the microphone also is sensitive to input at 90° and 270°. The directivity pattern has a null at 180°. The hypercardioidal pattern is similar to the cardioidal pattern, although there is small level of sensitivity to input at 180°, with a null at approximately 135° and 225°.

Directional microphone systems provide the capability of pointing the main beam, or beams, at *desired sound sources*, and pointing the null, or nulls, at *undesired sound sources*. Unfortunately, the actual spatial distribution of desired and undesired sound sources rarely conforms precisely to the directivity pattern of a given microphone. In fact, because of the complicated way in which noise reverberates in a room, noise from even a single source will impinge on the microphone from all directions. In an ideal environment, acoustic dampening materials can eliminate most reverberation, and directional microphones can be positioned to aim main beams at participants and nulls at undesired noises. However, typical home or office environments are reverberant, the users move about as they speak, and it is usually not practical to

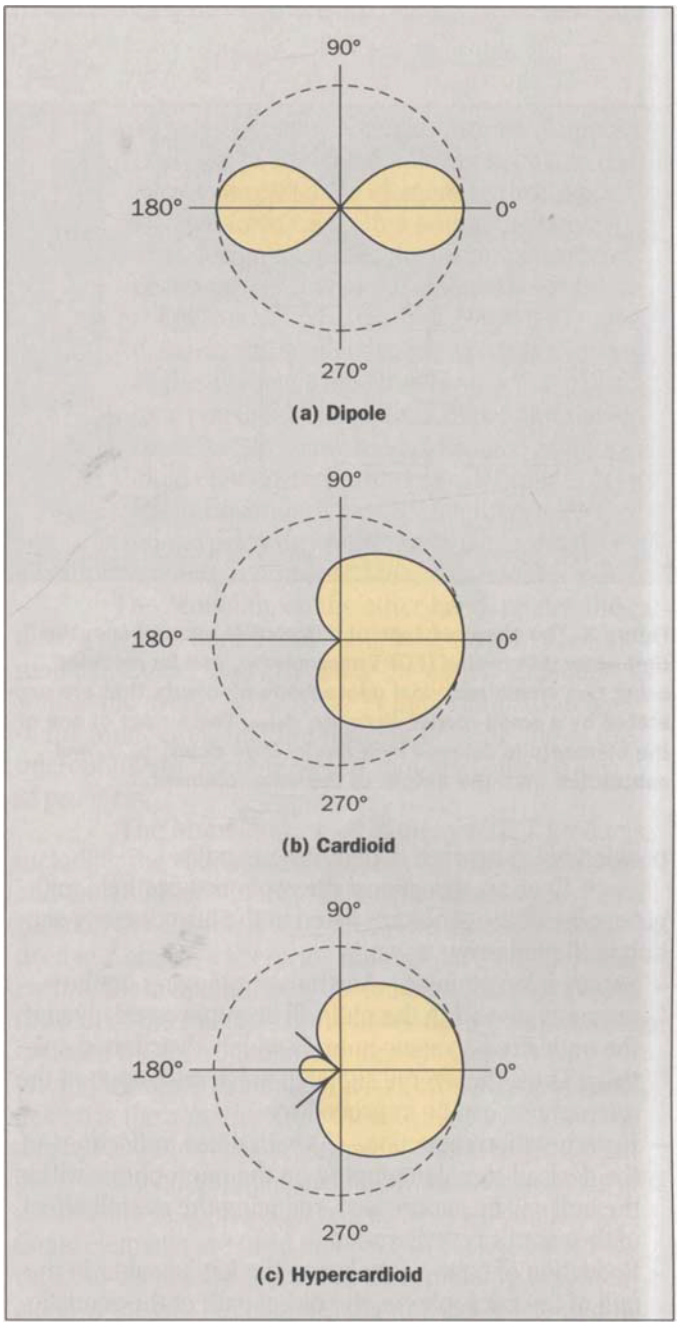

Figure 2. The directivity patterns of three directional microphones are shown. These are polar plots of the electrical output versus the angle of incidence for an acoustic input of constant amplitude and frequency. The plots are normalized with respect to the largest output. Figure 2A shows a *dipole* pattern, Figure 2b shows a *cardioid* pattern, and Figure 2c shows a *hypercardioid* pattern.

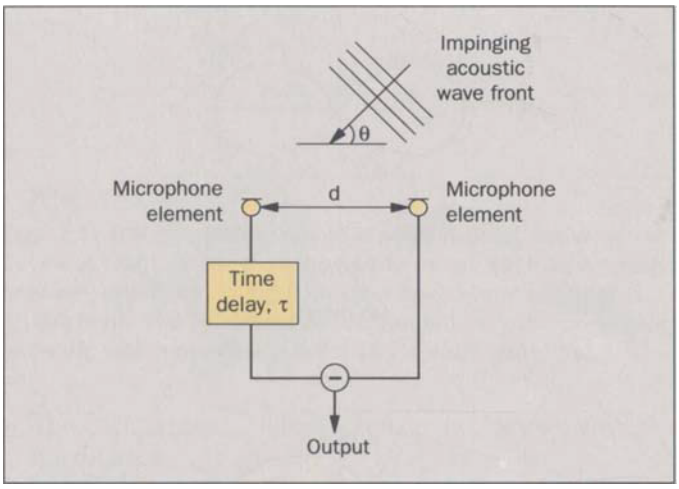

Figure 3. The simplest type of differential microphone, the *first order differential* (FOD) microphone, can be modeled using two omnidirectional microphone elements that are separated by a small *model distance*, d_{mod}. The output of one of the elements is delayed by a *model time delay*, τ_{mod}, and subtracted from the output of the other element.

position noise sources in microphone nulls.

Even so, directional microphones can help mitigate some of the problems listed in the introductory section in the following ways:

- Suppression of noise—Any noise impinging on the microphone within the null will be suppressed, even if the undesired acoustic noise is widely distributed spatially. Thus, the overall signal-to-noise ratio (SNR) of the microphone can be improved.
- Reverberation reduction—As with noise, reflections of the desired signal impinging on the microphone within the null will be suppressed, reducing the overall effect of the room's reverberation.
- Reduction of echo—By placing the loudspeaker in the null of the microphone, the direct path of the acoustic coupling between the two transducers can be essentially eliminated, reducing the acoustic echo of the speakerphone.

Directional microphones can be made in different ways. Conceptually, a useful way to create them is from a linear combination of low-cost omnidirectional microphones, often called *elements*, arrayed such that they s*patially sample* the acoustic environment. The two most basic types of directional microphones are *differential* (or *gradient*) microphones and *summing array* microphones. As their names imply, differential microphones use the difference of the elements' outputs to realize directivity, while summing arrays use the sum of the elements' outputs. There are advantages and disadvantages to each approach.

Differential Microphones. The simplest type of differential microphone is the *first-order differential* (FOD) microphone. A useful model of the FOD microphone is shown in Figure 3. Here, two omnidirectional microphone elements are separated by a small *model distance*, d_{mod}. The output of one of the elements is delayed by a *model time delay*, τ_{mod}, and subtracted from the output of the other microphone. While FOD microphones are seldom realized this way, the model is useful in that d_{mod} and τ_{mod} can be used to completely describe the system output at a given frequency, ω, of the acoustic input signal.

Specifically, assuming that d_{mod} is much smaller than the smallest wavelength, λ_{min}, in the acoustic signal, a good approximation to the amplitude of the microphone system output, $E_1(\omega, \theta)$, is:

$$E_1(\omega, \theta) \approx P_o \omega \left[\tau_{mod} + \frac{d_{mod}}{c} \cos\theta\right] \qquad \text{Equation 1}$$

where c is the speed of sound in air, θ is the angle of incidence of the acoustic signal, and P_o is a constant proportional to the amplitude of the acoustic wave.

The directivity patterns of Figure 2 can be realized by simply varying the relationship between τ_{mod} and d_{mod}. By setting τ_{mod} to 0, d_{mod}/c, and $d_{mod}/3c$, directivity patterns of the *dipole*, the *cardioid*, and the *hypercardioid*, respectively, may be realized.

By replacing each omnidirectional element of Figure 3 with a first-order element, each pointing in the same direction, a *second-order differential* or SOD microphone can be formed.[1]

If the distance between the SOD's elements, d_{sod}, is equal to the element's own model distance, and the delay after one of the SOD's elements, τ_{SOD}, is equal to the element's model delay, then the SOD's amplitude function $E_2(\omega, \theta)$ is essentially:

$$E_2(\omega, \theta) \approx P_o \omega^2 \left[\tau_{SOD} + \frac{d_{SOD}}{c} \cos\theta\right]^2 \qquad \text{Equation 2}$$

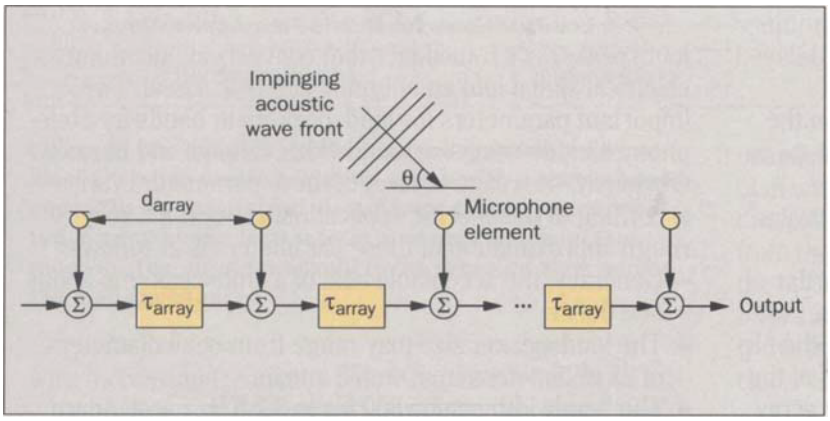

Figure 4. Summing arrays form their directivity patterns by *adding* the delayed outputs of two or more microphone elements. Each element is separated by a distance, d_{array}. The output of each element is combined with the output of the previous element's adder, delayed by θ_{array}. Parallel acoustic wave fronts impinging on the array from the angle θ add constructively, when τ_{array} is set appropriately.

From Equations 1 and 2 it can be seen that one drawback of differential microphones is that their output amplitude, that is, their sensitivity, is directly proportional to the frequency, or squared frequency, of the audio input. The FOD and SOD's responses decrease by 6 dB and 12 dB per octave with decreasing frequency, respectively, effectively forming a simple high-pass filter on the acoustic signal. This effect can be corrected by following the microphone with a complementary low-pass filter making the overall response flat.

One disadvantage of the multiple element approach to forming differential microphones is that electrical noise generated thermally in each of the elements does not share the high pass characteristic of the acoustic signal at the differential microphone's output. The reason for this is that the noise signal from each element is statistically independent from the other, unlike the acoustic signal. Thus, the signal-to-electrical noise (SNR_e) ratio at the differential microphone output can be low at low frequencies.

AT&T has developed some innovative microphone array designs based on the differential approach. One is the *Monolith* microphone. This is a FOD microphone with a hypercardioid directivity pattern. It is constructed by using a single cardioid element enclosed in a resilient housing such that the front and rear of the element are acoustically isolated from each other within the housing. This design increases the Monolith's model distance over that of the element's such that the relationship between the Monolith's model distance and model delay—which is the same as the element's model delay—is that of a hypercardioid. Normally, a hypercardioid element has the same model distance as the cardioid element, but with a smaller model delay. From Equation 1, we see that this smaller model delay *decreases* the hypercardioid's overall output amplitude, compared to the cardioid.

The Monolith, on the other hand, retains the cardioid's original model delay and, instead, *increases* the model distance, thus *increasing* the output amplitude, as well as the overall SNR_e. Another important advantage of the Monolith is that the resilient housing allows the microphone to be conveniently mounted on a variety of products.

The Monolith is used in many AT&T products. including the Videophone 2500, the Speakerphone 870, and the TalkBak™ loudspeaker/microphone peripheral (used in the Globalyst™ 360TPC). The Monolith also is used in Compaq's Presario* multimedia computers. In each of these applications, the loudspeaker is placed in the null of the microphone to lower the acoustic coupling between the microphone and loudspeaker.

Another example of an innovative microphone design is the *adaptive differential microphone.* Here the desired sound source is assumed to be in the front half-plane of the microphone and an unwanted noise source is assumed to be in the back half-plane. Two omnidirectional elements are used to form two back-to-back FOD cardioid beams. By appropriately combining weighted sums of the two FOD signals, the null can be steered to any desired direction in the rear half-plane. By optimally selecting the FOD signal weights to minimize the output energy of the system, the null is adaptively steered to the dominant source of unwanted noise.

Summing Arrays. Summing arrays form their directivity patterns by *adding* the delayed outputs of two or more microphone elements. Figure 4 shows such a linear summing array. Each element is separated by a

distance, d_{array}. The output of each element is combined with the output of the previous element's adder, delayed by τ_{array}.

Parallel acoustic wave fronts impinging on the array from the angle θ add constructively, when τ_{array} is set appropriately. Since, in summing arrays, all the acoustic signals are being added, they generally have a higher SNR_e than differential systems.

The total array length should be on the order of the wavelength of the lowest frequency to be preserved in the acoustic signal, while the distance between the elements should be smaller than half the wavelength of the highest frequency. Thus, for example, a 1.1 meter array of about 28 elements with an element separation of 4 centimeters (cm) would cover the standard 300-to-3300 Hz telephone bandwidth.

As the wavelength of the impinging signal approaches the length of the array, the differences in the pressure phase between consecutive elements becomes smaller and smaller. As a result, the array begins to lose its directivity, that is, the beam *broadens*. Hence, the array of Figure 4 can have a very narrow beam at high frequencies and a very broad one at low frequencies.

The result is that sound becomes more low-pass filtered as the angle from the center of the main beam increases. Accordingly, as a person in the process of speaking walks out of the beam, the speaker's voice sounds more and more muffled to listeners at the other end of the line. It is more desirable for all frequencies of the speaker's voice to be uniformly attenuated when leaving the beam. This can be accomplished by processing the elements' signals in subbands.

In subband processing, there are several arrays, each with different lengths and element spacings, designed to cover a specific band of frequencies with the same beam direction and width. The output of each array is band-pass filtered, accordingly, and the result is added to form the final output. Elements can be shared among the arrays as dictated by the individual arrays.

The advantages of summing arrays include higher SNR_e and the ability to form a very narrow beam. The main disadvantage is cost, since a large number of elements and their associated electronics are required. The large physical size of summing arrays is also a problem for desktop applications, though they can be effectively used in conference rooms and auditoriums.

Loudspeakers for Hands-Free Applications. A loudspeaker is a transducer that converts an incoming electrical signal into an outgoing acoustic signal. The important parameters for loudspeakers in hands-free telephony include cost, size, bandwidth, fidelity, and linearity. Obviously, the range of each of these parameters varies according to the specific application. However, a very rough approximation of these parameters is as follows:

- Generally, the acceptable cost of a loudspeaker is about one dollar;
- The loudspeaker size may range from cone diameters of 25 millimeters (mm) to 75 mm;
- The bandwidth can be 300 Hz to 3.2 KHz for standard telephone lines or 50 Hz to 7 KHz for commentary grade lines;
- The amplitude response generally varies by about 4 dB over the bandwidth; and,
- As will be discussed in the next section, the loudspeaker should never *clip*, that is, flatten out the signal peaks, or otherwise non-linearly distort the signal.

A loudspeaker consists of two main parts, the *driver* and the *enclosure*. The salient features of a driver are its diameter and maximum excursion limit, that is, how far the speaker diaphragm can move back and forth. Together, these determine the magnitude of sound the loudspeaker can faithfully, that is, linearly, produce.

The driver can be modeled as a simple spring and mass system. The input to the system is a force—proportional to the output speech signal—applied to the mass. The system's output power is proportional to the square of the mass's volume velocity times the acoustic radiation resistance of the air at the loudspeaker aperture.

This system has two important properties:

1. The acoustic radiation resistance is proportional to frequency squared, and
2. The system's mass volume velocity has a simple pole with a resonant frequency at

$$\omega_R = \sqrt{k/m}$$

where k is the stiffness coefficient of the spring and m is the mass.

Well below ω_R, the volume velocity is influenced by the stiffness and is, thus, proportional to frequency. Well above ω_R, in the mass controlled region, volume velocity is proportional to inverse frequency. Thus, it follows that the output power increases at 12 dB/octave

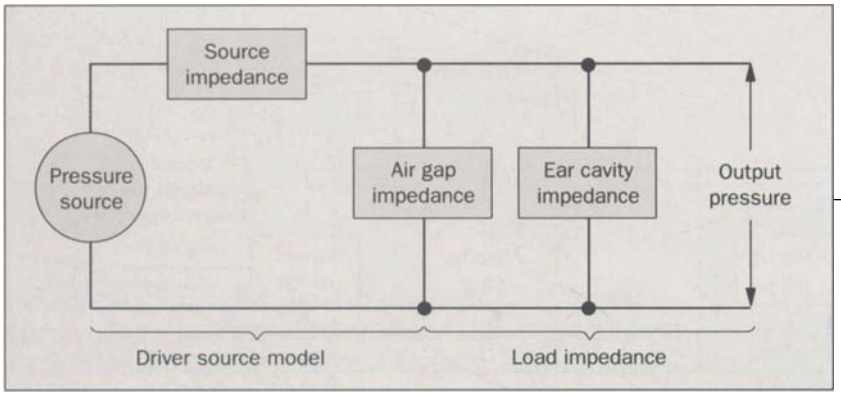

Figure 5. The handset receiver can be represented by an ideal driving pressure source in series with a source impedance. The ear cavity and air-gap impedances are connected in parallel and form a total load impedance on the receiver. The output pressure is the pressure seen across the load impedance.

with increasing frequency below ω_R, and is flat above ω_R.

Speaker drivers are natural dipoles. Recall that Equation 1 with $\tau = 0$ describes the directivity and sensitivity of a dipole microphone. It also describes the sound field generated by a dipole loudspeaker at low frequencies.

As with the dipole microphone, the dipole speaker's output is proportional to frequency, so the output level is small at low frequencies. For an observer to see an overall flat spectral response above the speaker driver's resonant frequency, this dipole effect must be addressed. This can be done by placing the driver in a sealed enclosure to absorb the output of one of the poles, creating an omnidirectional speaker with a much larger output level at low frequencies.

It is important that the volume of the enclosure, also called the *back cavity*, be large enough that the stiffness of the air within the enclosure does not overwhelm the inherent stiffness of the speaker driver, thereby raising the resonant frequency of the equivalent spring mass system and, thus, the low-end cutoff frequency of the loudspeaker.

To extend low frequency, or bass, response when using a small driver, the enclosure may be *vented* to the room, forming a *bass-reflex* system.[2] At and above the resonant frequency of the vent, the radiated sound will add constructively to the driver's sound, boosting the bass response.

Differential Microphones in Handsets. In telephone handset applications, it is more desirable for microphones to be selective with respect to the proximity rather than the direction of the acoustic source. Such devices are called *close-talking* microphones.[3] They are based on the differences between the near-field and far-field sounds impinging on the microphone.

So far, differential microphones have been discussed from the standpoint of the far field, where the distance between the microphone elements is small, as compared to the distance of the microphone from the sound source. For near-field sounds, the two distances are of the same order of magnitude. This is important because acoustic energy dissipation is proportional to the *square* of the distance from the source. That is, 6 dB of energy is lost each time the distance from the source is doubled. If the two elements are far from the source, then each sees approximately the same energy, but if they are close to the source, the closer element sees more energy than the farther element.

Thus, for example, if the differential microphone of Figure 3 is placed in the near field of a talker such that the talker is in the main beam, the element closer to the talker will have a stronger input signal than the one farther away. This effect changes the output amplitude function of the microphone from that of Equation 1. The on-axis ($\theta = 0$) output for a FOD for the case when $\tau = 0$ is:

$$E_1(\omega, r) \approx \frac{P_0 d}{r^2}\sqrt{1 + \frac{\omega^2}{c^2}}\, r^2 \qquad \text{Equation 3}$$

where r is the distance of the center of the elements from the source.

From Equation 3, it is clear that when $\omega << c/r$, the output magnitude becomes independent of the frequency of the nearby source. On the other hand, when $\omega >> c/r$, the output magnitude is directly proportional to ω.

Equation 3 also shows that as r increases, the frequency dependency becomes important at lower and lower frequencies. Hence, near-field sources are passed at low frequencies, while far-field sources are attenuated, that is, the FOD is *near-field selective* at low frequencies. The effect is even more pronounced for SODs, since they have frequency independent terms in the near field as well, but have even greater rejection of far-field source low frequencies.

The overall gain and the cut-off frequency for the near-field frequency independence is a function of r. Hence, if a user varies the position of his or her mouth relative to the microphone and changes r, the speaker will dynamically change the gain and filtering characteristics of the overall system. This can be annoying to the user at the other end. If r were known, different gain and low-pass correction filters could be applied as the user changes position. In fact, r can be estimated by compar-

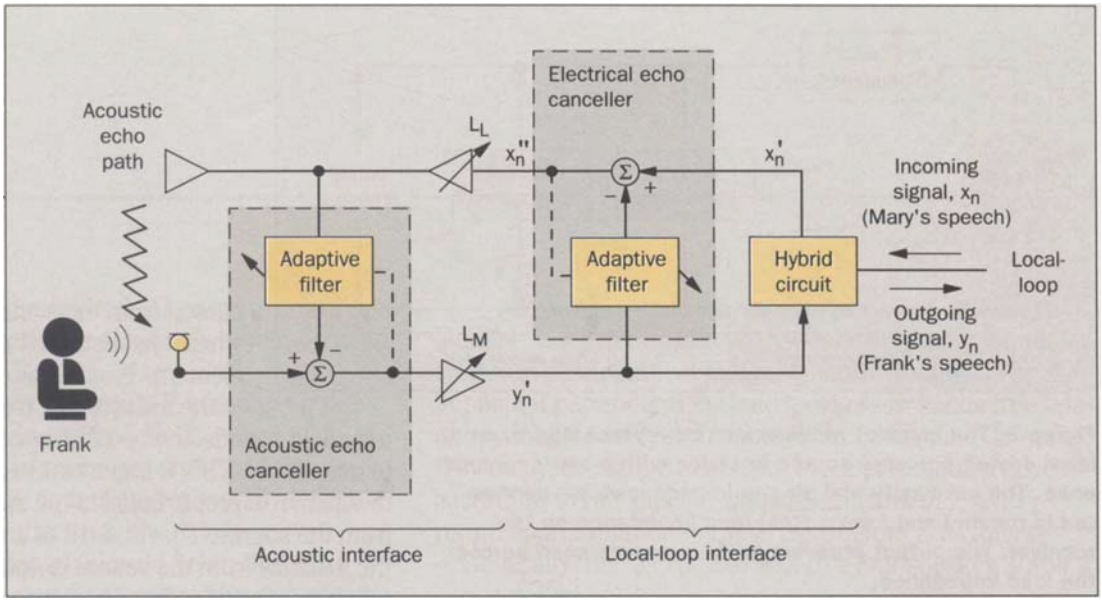

Figure 6. This simplified block diagram of an ideal echo canceling speakerphone is divided into two parts: the local loop interface and the acoustic interface.

ing short-time energies at the output of each of the microphone elements. The larger the nearer element's energy is compared to that of the farther element, the closer the microphone is to the user. This measurement can be used to select appropriate gain and low-pass filters for each measured distance, counteracting the unwanted modulation due to user movements.

Close-talking microphones also benefit significantly from the noise-reducing properties of their directivity pattern, as was pointed out in the discussion on differential microphones.

Receivers (Speakers) in Handsets. In handset applications, the electrical-to-acoustic transducer is commonly called a receiver, rather than a loudspeaker. A common problem in handset receiver design is the *acoustic leakage* of signals at low frequencies when the ear-piece of the handset is not fully pressed against the user's ear to create an acoustic seal.

An equivalent circuit of the receiver/air-gap/ear system is shown in Figure 5. The receiver is represented by an ideal driving pressure source in series with a source impedance. The ear cavity and air-gap impedances are connected in parallel and form a total load impedance on the receiver. The output pressure is the pressure seen across the load impedance. The source impedance of the handset receiver is normally designed for the case where the gap leakage is negligible—when the handset is sealed tightly to the ear. In that case, the output pressure is about half the driving pressure.

Unfortunately, when there is a leak in the acoustic seal, the air-gap impedance becomes very small at low frequencies and makes the *ear cavity* impedance appear negligibly high. In this case, most of the pressure drop will be seen across the *source* impedance, rather than the total load. To correct this situation, a *low acoustic impedance receiver* (LAIR), which has a source impedance more comparable to the leakage impedance at low frequencies, has been designed.

Digital Signal Processors for Acoustics

In this section, the focus is shifted from transducers to DSP techniques for acoustic processing. Progress made in the following areas is reviewed:

- *Echo cancellation and control*, which provides full-duplex hands-free audio communications.
- *Speech enhancement*, using single or multiple microphones, which reduces ambient noise and improves the quality of transmitted speech and the performance of voice processing systems.
- *Active noise cancellation* (ANC), which reduces ambient noise at the ear-piece of telephone handsets and improves the quality of the received speech.

An Ideal Speakerphone. Figure 6 shows a simplified block diagram of an ideal echo canceling speakerphone. It is divided into two parts: the *local loop interface* and the *acoustic interface*. The key feature of the local loop is that it is a bidirectional channel—that is, both the outgoing (y'_n) and incoming (x_n) signals use the same pair of wires.

A hybrid, which is a relatively simple, low-cost circuit, is used to separate the outgoing from the incoming signal at the local-loop interface. Usually it is only partially successful and reduces the contribution of (y'_n) in its output, (x'_n), by only a few dB. The *electrical echo canceler* (EEC)[4,5] attempts to remove the remaining echo signal, by mimicking the behavior of the combination of the hybrid and local-

loop circuits, that is, the *electrical echo path* (EEP). The EEC then produces an *electrical echo estimate* and subtracts it from the actual electrical echo and incoming signal, producing y''_n. The EEC accomplishes this via an *adaptive filter.*

The electrical echo path is modeled as a weighted combination of the current and past versions of the outgoing signal. The number of weighted samples, N_E, required to form a faithful model of the EEP is called the *electrical echo path length.* Whenever the outgoing signal is active (for example, when Frank is speaking) and the incoming signal consists of only low-level circuit noise (when Mary is silent), the adaptive filter adjusts its weights in an attempt to drive x''_n to zero. When this occurs and outgoing y'_n is sufficiently rich in spectral content, the adaptive filter's weights will be the same as the EEPs. The most common adaptive filtering algorithm is the *normalized least mean squares* (NLMS) method. NLMS's complexity is roughly $2N_E$ multiplications per sample period.

When incoming x_n is active, the adaptation is inhibited—we don't want the incoming signal eliminated—but the process of producing the electrical echo replica and subtracting it from the echo plus incoming signal (producing the echo-free result, x''_n) is continued.

The process of deciding whether the incoming signal is present or not is easy when the outgoing signal is a low-level background signal. All that needs to be done is to monitor the level of x'_n. However, when outgoing signal y'_n is active, the signal in x'_n can be either electrical echo or echo and incoming signal. The latter situation is called *double talk.* It is the job of a *double-talk detector* (not shown in Figure 6) to detect this situation and inhibit the filter's adaptation process. An explanation of an elementary double talk detector is given by Duttweiler.[5] This is generally the most difficult part of an echo canceler to design.

The local-loop interface of the speakerphone in Figure 6 may or may not be necessary, depending on the channel to which the phone is connected. For instance, in many videophone applications, the connection between the audio terminals is two separate uni-directional channels.

On the acoustic side of the speakerphone, an *acoustic echo canceler* (AEC), similar in principle to the EEC, is shown. There are a few important differences that make the AEC more difficult to design than the EEC:

- The acoustic echo path length is much longer than the electrical echo path length, namely, hundreds rather than tens of milliseconds.
- The acoustic echo path changes very quickly with the movement of any object or individual in the room, including the user.
- While there is a guarantee of at least some attenuation of the echo by the hybrid in the electrical echo path, there often is *gain* of 20 dB or more in the acoustic echo path because of the loudspeaker and microphone amplifiers. (This 20-dB gain assumes that the speakerphone loudspeaker volume estimate control is at its maximum setting and an omnidirectional microphone is used. A directional microphone can reduce this gain to about 10 dB.)

The effect of the longer echo path is twofold. First, the computational complexity of the adaptive filter (the number of multiplications) increases by an order of magnitude, increasing proportionally the cost of the adaptive filter's implementation.

Second, the convergence speed of the algorithm for adaptation is slower for longer filters, especially when the exciting signal is highly self-correlated, as is the case with speech. This slow convergence is unacceptable because the acoustic echo path changes so quickly.

The problem is compounded by the fact that the echo is so loud that any residual echo is likely to be heard and be annoying. While the progress of technology increasingly makes moot the issue of computational complexity, the slow convergence issue has required the use of new adaptive filtering algorithms, such as subband echo cancelation (SBECs).[6] This method :

- Divides the excitation and echo signals into M equal-sized subbands;
- Sub-samples them by a factor R (that is using only every R'th sample, discarding the rest), where R is close to, but a little less than M;
- Performs echo cancellation on each subband separately (using NLMS, for example); and then
- Upsamples and recombines the echo residuals into one full-band echo residual.

This technique lowers the average computational complexity by about a factor of $1/R$ because, in each subband, the effective echo path length (EPL) is a factor of $1/R$ smaller, and there is a factor of R more time to perform each computation, but there are $M \approx R$ subbands to compute. Convergence is also accelerated because the subband signals tend to be less self-correlated than the full-band signal.

The main disadvantage of SBECs is the delay incurred in the subband analysis and synthesis filter banks. This delay can be from 10 to 100 ms, depending on the number of subbands. Generally, the more subbands, the longer the delay.

At the very beginning of a call, or when the echo paths have changed between excitation talk-spurts, the echo cancelers have not had a chance to adapt to the echo path characteristics. Because the gain in the acoustic path is often greater than the loss in the electrical path, there can be an overall gain in the acoustic-hybrid loop. This is an unstable situation, since any input, say from the room, will be amplified again and again each time it passes through the loop. This situation is called *howling* or *singing*. The loudspeaker and microphone *switched-losses*, L_L and L_M, are used to prevent this from happening.

A *switched-loss controller* continually measures the amount of gain in each echo canceler-echo path combination to determine the amount of overall loss required to keep the system stable. The resulting loss is then divided between the loudspeaker and microphone switched-losses. The strategy is generally to switch most of the loss to the microphone circuit when only the incoming signal is active, most of the loss to the loudspeaker circuit when only the outgoing signal is active, and half in each when they are both active.

Another technique sometimes used for echo control is a *center-clipper* (not shown in Figure 6), which is usually placed after the echo canceler and switched-loss. In center-clipping, speech signals below a predetermined threshold are simply set to zero. Large amplitude signals, like the incoming signal on the EEC side, are passed, but small signals, like the residual echo, are eliminated.

Both switched-loss and center-clipping cause the annoying side effect of modulating the background noise level of the desired signal. As a result, noise is often purposely added after these functions to "fill" the signal that was removed.

Influences on Speakerphone Design. There are six main factors that affect the design of speakerphones:

- Speakerphone costs,
- Channel delay,
- Unidirectional/bidirectional channels,
- Loudspeaker/user/microphone geometry,
- Echo path linearity, and
- Room reverberation characteristics.

Each of these is discussed below with regard to how they affect two specific applications—desk-top/audio-only (DTAO) and audio for video-conferencing (AVC).

Cost. The cost issue loosely translates to available computational complexity and memory. The DTAO application is much more cost sensitive than the AVC application and, hence, cannot afford the same resources. Although, as technology improves, the cost of processing and memory will decrease for both applications.

Channel Delay. The channel connecting the speakerphones affects their design in two ways. The first is channel delay. The perception of echo is a function of the signal's round-trip-delay (RTD) (2D in Figure 1) and the degree which the echo is attenuated in the echo path. The mean required attenuation to eliminate the perception of echo increases with RTD. When the RTD is 20 ms, the required attenuation is only about 10 dB, but when the RTD reaches about 800 ms, the echo must be almost completely eliminated. Typical channel delays are 10 ms for DTAO and 400 ms for AVC. The DTAO delay is short, since it generally uses standard telephone connections that have been designed over the years for small delay. The AVC delay is long because the delay for video coding is large, and the audio must be delayed to synchronize with the video.

As a result, DTAOs only need to attenuate echo to about 10 dB below the incoming level, while AVCs must remove it altogether. This means that the AVC must use all the available means described above to suppress echo—namely, a long acoustic SBEC with moderate amounts of switched-loss, center-clipping, and noise-fill.

Taking into account the 20-dB gain in the acoustic echo path (AEP) mentioned above, DTAOs need to attenuate the echo by about 30 dB. This can be accomplished using switched-loss techniques alone, or by using switched-loss combined with a short echo canceler having an EPL of about 50 ms. In the former case, the resulting unit is not truly full-duplex, but it is affordable and users are still able to interrupt each other effectively. In the latter case, near full-duplex operation can be achieved at a higher cost.

Bidirectional/Unidirectional Channel. The other important channel characteristic is bidirectional versus unidirectional channel interface. The DTAOs are usually connected to the bidirectional local loop, while AVCs often connect to each other via unidirectional channels.

Loudspeaker/User/Microphone Geometry. The geometry of the loudspeaker, user, and microphone—as in Frank's room in Figure 1—is critical in a number of ways. The volume setting the user selects for the loudspeaker is proportional to the distance between them. The overall gain in the hybrid-acoustic loop varies directly with this gain setting and as the inverse square of the distance between the microphone and loudspeaker. Finally, the signal-to-background room-noise ratio is inversely proportional to the distance between the user and the microphone.

DTAOs have microphone-user, user-microphone, and loudspeaker-microphone distances of about 1, 1, and 0.1 meters, respectively; for AVCs, all of these distances are between one and several meters.

Echo Path Linearity. Because the adaptive filters in the echo cancelers form their echo estimates as weighted (or *linear*) combinations of past and present excitation samples, the echo paths must also behave in the same way. That is, the echo paths must be linear. For the acoustic path, this means that the loudspeaker must not clip the incoming signal, and the loudspeaker and housing of the speakerphone must not "buzz" or "rattle." If these conditions occur, the AEC simply will not work and the speakerphone will most likely go unstable and begin to howl.

Room Reverberation Characteristics. The room reverberation characteristics determine the shape and length of the AEP. Rooms with reflective walls, floors, and ceilings, like kitchens, have longer reverberation times than rooms with absorbing surfaces. Reverberant rooms tend to have higher background noise levels, since noise energy persists longer. Typically, the magnitude of the impulse response of a room has an exponential decay, and low frequencies tend to persist longer than high frequencies.

It is only recently that echo cancelers have begun to appear in DTAOs. In fact, most units currently deployed do not have them. This is because until the advent of SBECs about five years ago, the AEC problem was too hard to solve, and the cost of implementing the adaptive filters, especially the AEC, was too expensive. In speakerphones without echo cancelers, usually the only approach undertaken for echo control was switched-loss. To keep the system stable without the aid of ECs, the combined losses needed to be 40 dB or more under worse-case conditions. These are truly half-duplex systems.

An example of a DTAO is AT&T's Speakerphone 870. It uses an EEC to prevent singing in the hybrid-acoustic loop, allowing less switched-loss and more full-duplex-like performance. Examples of the AVC include AT&T's QuiteQuiet™ Acoustic Echo Canceler and the AEC for the AT&T Vistium™ Personal Video System 1200 and 1300.

Speech Enhancement. Speech enhancement involves processing noisy speech signals to make the noisy signals less objectionable to listen to for long periods of time. Speech enhancement is also used to improve the performance of voice processing systems, such as speech recognizers and coders. The problem of enhancing speech corrupted by ambient noise can be characterized by the type of noise, the number of microphone outputs available for processing, and the nature of the voice communication system.

The interfering noise may be background sounds produced by street traffic, competitive talkers, room reverberation, and office equipment, such as fans and disk drives. Furthermore, the noise may be narrowband (that is, tonal), broadband, or a combination of the two; stationary or non-stationary; and statistically correlated or uncorrelated with the desired signal.

When an interfering noise source is uncorrelated with the desired speech signal and can be sensed, relative to the speech, by one or more secondary microphones, then adaptive cancellation techniques can be used to effectively remove the noise from the primary microphone.[7] Alternatively, if only a single output is available containing the noisy speech, then one is faced with the more formidable problem of enhancing the speech without the benefit of an explicit noise reference signal. A number of approaches for addressing this problem can be found in the literature.[8] Generally, the approaches fall into one of three categories: *spectral subtraction of interfering noise, speech harmonic enhancement, and model-based enhancement by resynthesis.*

Spectral Subtraction. The basic principle of *spectral subtraction* is the concept that "correcting" the spectral magnitude of noisy speech is perceptually more important than correcting the phase. The short time spectral magnitude of the noisy speech signal is computed, from which is subtracted an estimate of the noise spectral

magnitude made during periods when speech is not present. If the result after subtraction is less than zero, then the spectral magnitude is set equal to zero.

The residual spectral magnitude is then combined with the original phase and transformed back to a time waveform. The method of spectral subtraction is applicable only to stationary noise with a SNR of at least 6 dB, since it assumes that periods when speech is not present can be identified and that the noise estimates made during these periods are representative of noise during periods of speech.

In addition, because of the zeroing of negative values caused by magnitude subtraction, the residual spectrum contains random tone bursts, which are often referred to as "musical noise" because of their perceived sound. Much work in the area of spectral subtraction has been directed at minimizing the musical noise, while maintaining the overall quality of the enhanced speech.

Harmonic Enhancements. The *harmonic enhancement* method capitalizes on the fact that voiced sounds are periodic and that, by estimating the pitch period accurately, a comb filter can be used to pass the harmonics of speech but reject the frequency components between the harmonics. Comb filtering can be used to suppress a range of background noises, and it has been used frequently as a means to combat interference from a competing talker.[9]

The major difficulty in the case of an unwanted talker is accurately tracking the preferred talker's fundamental frequency in the presence of the speech from the unwanted talker. Generally, comb filtering, even with accurate pitch information, tends to decrease the intelligibility at various SNRs, although the processed speech sounds less noisy because the SNR is improved. Nevertheless, the harmonic enhancement method has had some success in automatic speech recognition when the interference is another voice.[10]

Model-Based Speech Enhancement. *Model-based speech enhancement* exploits the underlying model for speech production by estimating the model parameters from the noisy speech, and then generating enhanced speech by a synthesis system based on the same underlying speech model. The model parameters, typically, consist of the pitch period and the coefficients of the filter used to represent the vocal tract.

Several model-based enhancement systems have been developed with the speech model parameters estimated by what is called the *maximum likelihood* (ML) method that accounts for the presence of background noise. Subjective listening tests indicate that although the quality of the speech is improved with this method, the improvement in intelligibility is not clear.[11]

In a more recent study,[12] speech enhancement was investigated using hidden Markov modeling (HMM) of the speech and noise signals. Tests with additive white noise at 10 dB input SNR showed significant reduction in the input noise, but with some of the speech accompanied by a low-level noise that made some of the words sound hoarse.

Commercial products and systems that enhance the voice quality by suppressing background noise are being introduced. For example, when AT&T's QuiteQuiet AEC is introduced this year, it will have an optional speech enhancement feature to reduce stationary background noises that are typical of office environments.

Active Noise Cancellation. Active noise cancellation systems generate a canceling sound field or "anti-noise" that destructively interferes with undesired acoustic noise to produce a zone of silence over a region of space. The development of microelectronics technology, especially DSP hardware and miniature transducers, has advanced this decade's old technology sufficiently to make it commercially viable.

For example, ANC can now be found in commercial communication and hearing protection headsets that are effective for canceling low-frequency noise below 1 KHz, together with passive means effective above 1 KHz. Most of these commercial headset systems use analog feedback control in which the residual, that is, the canceled noise field within the ear-piece, is measured by an error microphone whose output is filtered and fed back to the ear-piece transducer, usually a loudspeaker. The transducer, in turn, produces the canceling sound field at the microphone.

Performance is limited by phase shift in the closed loop, which, due to stability considerations, restricts optimal feedback gain levels. The dominant phase shift occurs in the loudspeaker and its associated acoustic cavities due to multiple coupled resonances. Consequently, the maximum cancellation achieved is about 15 dB, typically peaking somewhere between 300 to 700 Hz.

An even greater challenge is presented in applying ANC to telephone handsets, where the transfer function of the receiver/microphone varies with the positioning of the ear-piece about the ear.

One solution monitors feedback variability and adjusts the gain accordingly to maintain performance. Alternatively, adaptive feed-forward control can be used where an additional microphone, placed outside the ear-piece cavity, produces a *reference* signal that is fed into an adaptive finite impulse response (FIR) filter controlled by the feedback signal to produce the canceling drive signal for the receiver. The FIR coefficients must be adapted with the "Filtered-X" algorithm,[13] which takes into account the presence of the receiver/microphone transfer function.

The need for high correlation between the two microphone outputs requires they be placed physically close. Again, the receiver's inherent delay limits performance to lower frequencies, where speech energy, but not intelligibility, is dominant.

To judge consumer's reactions to an ANC ear-piece, prototype cellular and public telephone handsets were evaluated in a focus group study. Most users found the technology to be attractive, claiming that ear-piece noise reduction makes voices seem clearer and fuller.

Conclusions

This paper has explained some of the ways that ongoing advances in electroacoustics and DSP are improving the audio quality of AT&T's terminal equipment products. Telephones and speakerphones benefit from the application of both disciplines. For example, acoustic echo cancellation simply will not work if poor speakers and housings are selected.

Directional microphones, together with echo cancelers, switched-loss, and center-clippers, all contribute to the elimination of echo and instability. Examples of future work include making progress in providing:

- Low-complexity, fast-converging, delayless adaptive filters for echo cancelers;
- Faster, more reliable, double-talk detectors;
- Small, broadband, low-cost loudspeakers and microphones; and
- Artifact-free speech enhancement systems for noisy environments.

Progress along these fronts will further enhance the high quality of AT&T's terminal equipment and continue to ensure that AT&T is the vendor of choice for voice products and services.

Acknowledgments

The authors would like to thank Drs. L. R. Rabiner and G. W. Elko for their assistance in the preparation of this paper.

(Manuscript approved February, 1995)

*Trademarks

Presario is a registered trademark of Compaq Corporation

References

1. G. M. Sessler, J. E. West and R. A. Kubli, "Unidirectional, second-order gradient microphone," *Journal of the Acoustical Society of America*, Vol. 86, 1989, pp. 2063-2066.
2. R. H. Small, "Vented Box Loudspeaker Systems," *Journal of Audio Engineering Society*, Vol. 21, 1973.
3. J. E. West, G. W. Elko, D. R. Morgan, and R. A. Kubli, "Position-tolerant differential microphones for noise environments," AT&T Bell Laboratories Technical Memorandum: 11227-901228-18TM.
4. Sondhi, M., Presti, A. J., "A Self-Adaptive Echo Canceler," *Bell System Technical Journal*, Vol. 45, 1966, pp. 1850-1854.
5. Duttweiler, D. L., "A Twelve Channel Digital Echo Canceler," *IEEE Transactions on Communications*, Vol. 26, No. 5, May 1978.
6. W. Kellermann, "Analysis and Design of Multirate Systems for Cancellation of Acoustical Echos," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 1988.
7. B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Inc., 1985.
8 J. S. LiMary (Editor), *Speech Enhancement*, Prentice-Hall, Inc., 1982.
9. T. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection," *Journal of the Acoustical Society of America*, Vol. 80, pp. 911-918.
10. M. Weintraub, "A Computational Model For Separating Two Simultaneous Talkers," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1986, pp. 81-84.
11. J. S. Lim, "Speech Enhancement," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1986, pp. 3135-3142.
12. Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proceedings of IEEE*, Vol. 80, 1992, pp. 1526-1555.
13. D. R. Morgan, "An analysis of multiple correlation cancellation loops with a filter in the auxiliary path," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, August 1980, pp. 454-467.

John C. Baumhauer, Jr., *is a distinguished member of technical staff (DMTS) in the Voice and Audio Processing Forward Looking Work Group at AT&T Bell Laboratories in Indianapolis, Indiana. He does mathematical simulation and optimization of systems and components to support the development of electroacoustic technology and products for AT&T Consumer Products Division. He joined the company in 1973. He has a B.S.M.E. degree from Drexel University in Philadelphia, Pennsylvania, and an M.S. degree and Ph.D. in mechanics from Rensselaer Polytechnic Institute, Troy, New York.*

Scott H. Early *is technical manager of the Voice and Audio Processing Forward Looking Work Group at AT&T Bell Laboratories in Indianapolis, Indiana. His group works on transducer and signal processing technology to support the next generation of terminals for AT&T Consumer Products Division. He joined the company in 1977. He has a B.S.E.E. degree from Purdue University in West Lafayette, Indiana, and an M.S.E.E. degree from the University of California at Berkeley.*

Steven L. Gay *is a member of technical staff in the Acoustics Research Department at AT&T Bell Laboratories in Murray Hill, New Jersey. His research interests include filtering and acoustic echo cancelation. He joined the company in 1980. He has a B.S.E.E. degree from the University of Missouri-Rolla, an M.S.E.E. degree from the California Institute of Technology-Pasadena, and a Ph.D. in electrical engineering from Rutgers University in Piscataway, New Jersey.*

John H. Fikus *is a technical manager in the Audio Display Terminals Development Department at Global Business Communications Systems (GBCS) in Middletown, New Jersey. His group is responsible for audio, acoustics, and software for business terminals. He joined the company in 1977. He has a B.S. degree in physics from Fairleigh Dickinson University, Teaneck, New Jersey; and an M.S. degree and Ph.D. in materials science from the University of Virginia in Charlottesville.*

Michael A. Zuniga *is technical manager of the Advanced Development Acoustics and Voice Processing Systems Group in the AT&T Custom Electronic Systems Business Unit in Arlington, Virginia. His group is responsible for the development of advanced noise suppression technologies, including active noise cancelation, noise canceling microphones, and speech enhancement algorithms. He joined the company in 1987. He has a B.S. degree from Pennsylvania State University at State College, an M.S. degree from Drexel University in Philadelphia, Pennsylvania; and a Ph.D. from the Massachusetts Institute of Technology in Cambridge, Massachusetts, all in physics.*