Telephone conversations in which the speakers do not use handsets or wear microphones should ideally be as convenient and intelligible as if participants were in the same room. But sometimes room acoustics interfere. Ways are being found to overcome such obstacles.

Seeking the Ideal in "Hands-Free" Telephony

David A. Berkley and Olga M. Mracek Mitchell

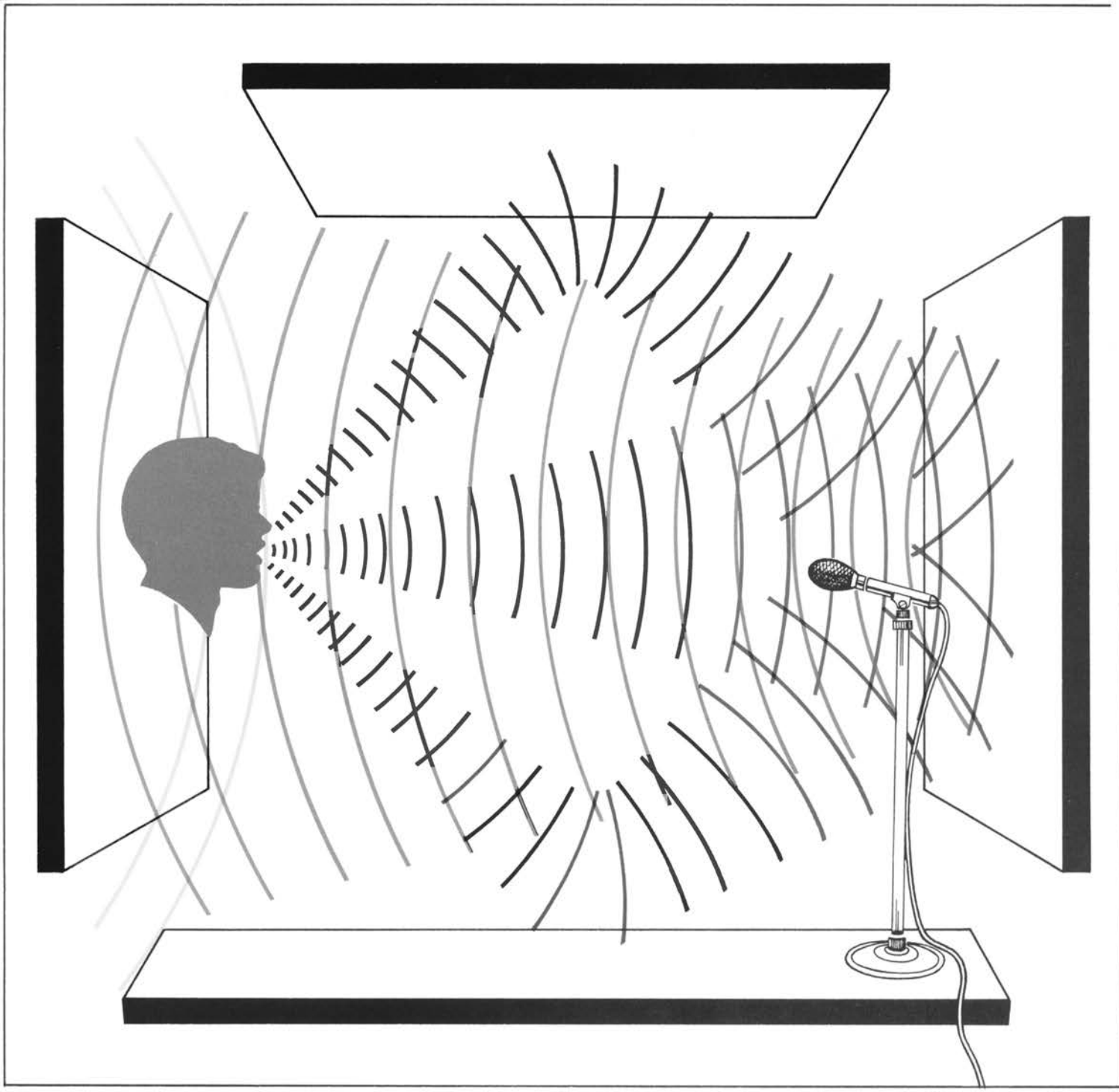
IDEALLY, CONVERSATION BETWEEN PEOPLE at different locations should be as easy and of as high quality as conversation across a table or desk. The components of the communication system should not unnecessarily restrain the participants—in other words, they should not be required to hold a handset or wear a microphone. Much progress has already been made toward such "hands-free" telephony (see the RECORD articles *Conferences and Classes Via PCT: If You Can't Come, Call,* April 1973, and *The 4A Speakerphone—a Hands-Down Winner,* September 1973). However, echoes and acoustic coupling can still degrade performance in some cases. In this article, we will review new approaches to these problems now being investigated at Bell Labs.

The hands-free telephony situation in which only one speaker is present at each end of the system is the one in which the problems are closest to resolution. (For a discussion of the more complicated situation in which there are groups of people at either end, see "The Cocktail Party Effect," page 322.) Speech picked up by a single microphone has a hollow and blurred sound as though the talker was speaking in a barrel. This distortion results from the microphone picking up not only the speech coming directly from the talker, but the reflections from the walls and furniture in the room as well (illustration on opposite page). These reverberation effects are not as disturbing for a binaural listener present in the room or a person listening binaurally at a remote location through suitable headphones.

Room reverberation consists of a series of echoes of the original speech, delayed from a few milliseconds to several seconds. The earliest echoes are single reflections from table tops and walls near the talker and the microphone. Later echoes are multiple reflections from the walls of the room which gradually become weaker; in a typical office their level is attenuated by 60 dB after about one-third second. This time is usually referred to as the reverberation time.

For speech sounds heard through a single microphone, early and late echoes seem to be perceived differently. The most important perceptual effect of early echoes is to change the frequency spectrum of the speech sounds, giving the speech a hollow quality. The effect is especially annoying in small, hard-walled rooms. Late reflections can be heard as distinct echoes of the speech sounds. This effect can be heard in an auditorium, in which

The accompanying record demonstrates the cures for reverberation effects described in this article. To remove the record, tear it along the perforated line. To store it, insert it in the slots inside the back cover.

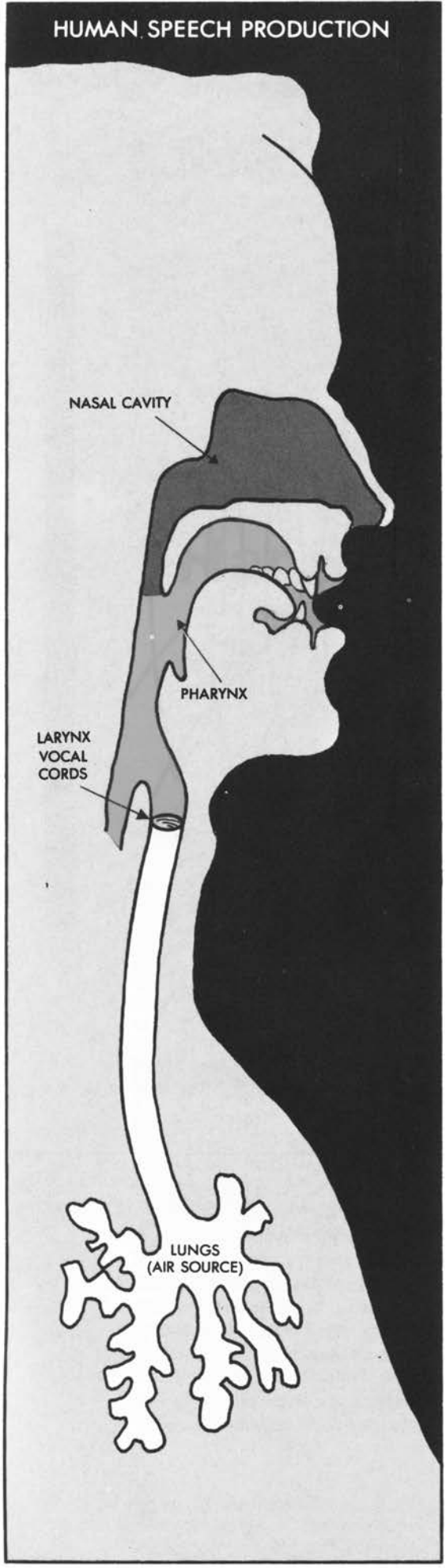


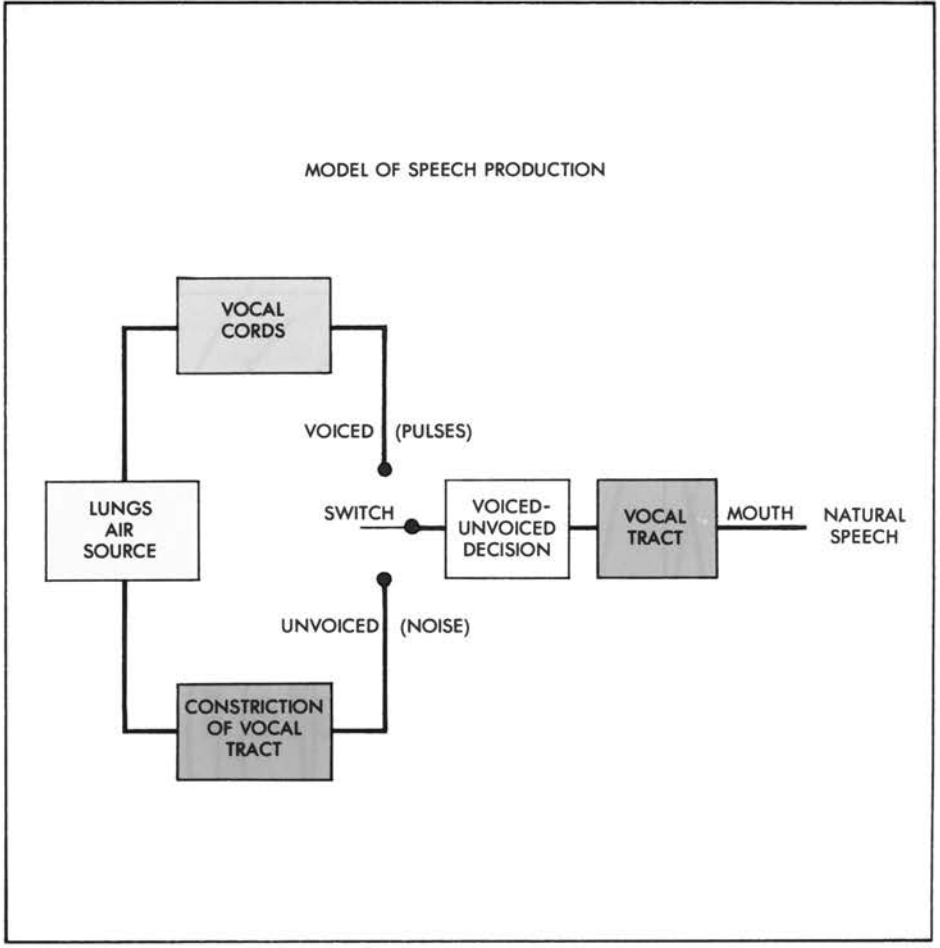
Reverberation in a room consists of echoes of the original speech. The earliest echoes are single reflections from table tops and walls near the talker and the microphone. Later echoes consist of multiple reflections from the wall surfaces of the room.

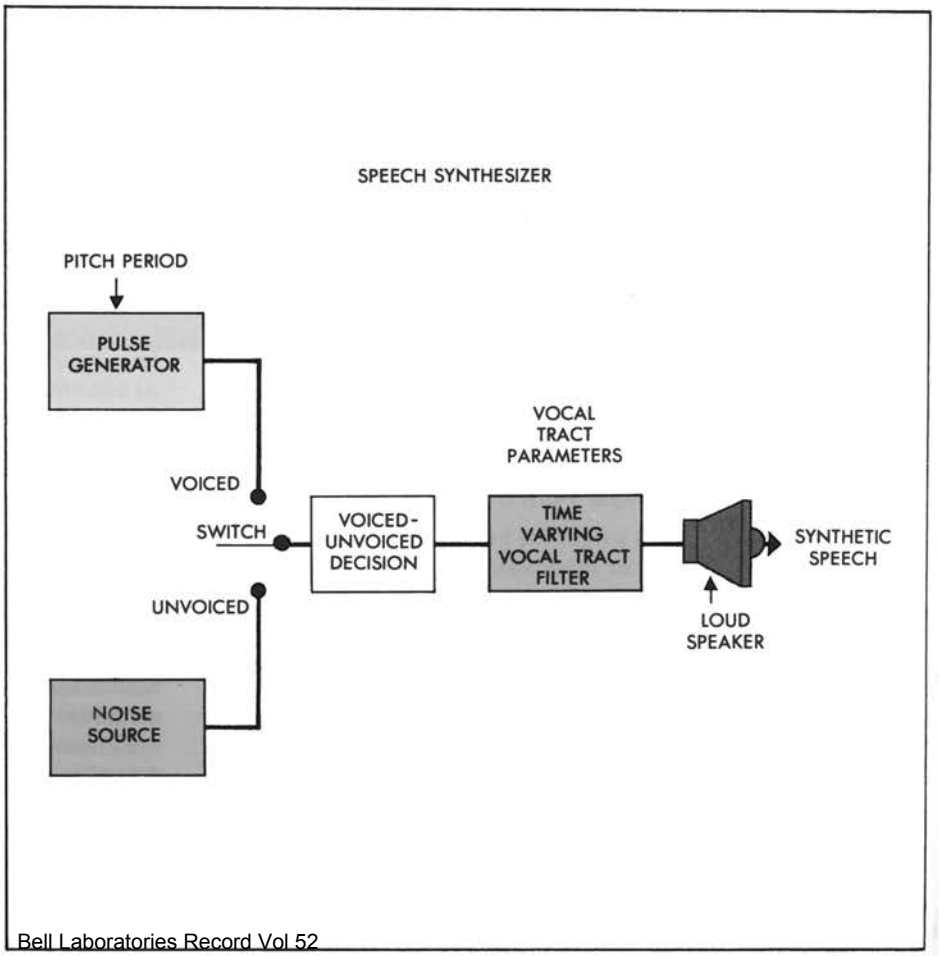
speech generally does not have the hollow sound caused by early echoes but is smeared out in time by later echoes.

Several methods of reducing the effects of reverberation in telephony have been studied. Essentially, these methods either (1) reduce the amplitude of the echoes reaching the microphone by treating the room with sound absorbing material or (2) remove the effect of the echoes by signal processing using one or more microphones. We will describe several of these signal processing methods.

Echoes with delays that are long relative to the duration of speech sounds form reverberant tails


HUMAN SPEECH PRODUCTION
NASAL CAVITY
PHARYNX
LARYNX
VOCAL
CORDS
LUNGS
(AIR SOURCE)


MODEL OF SPEECH PRODUCTION
VOCAL
CORDS
VOICED
(PULSES)
LUNGS
AIR
SOURCE
SWITCH
VOICED-
UNVOICED
DECISION
VOCAL
TRACT
MOUTH
NATURAL
SPEECH
UNVOICED
(NOISE)
CONSTRICTION
OF VOCAL
TRACT


SPEECH SYNTHESIZER
PITCH PERIOD
PULSE
GENERATOR
VOCAL
TRACT
PARAMETERS
VOICED
SWITCH
VOICED-
UNVOICED
DECISION
TIME
VARYING
VOCAL TRACT
FILTER
SYNTHETIC
SPEECH
LOUD
SPEAKER
UNVOICED
NOISE
SOURCE

HOW SPEECH IS PRODUCED

Speech is produced in the vocal tract, a resonant system consisting of the mouth and a tube called the pharynx, with the nasal cavity as a side branch (see diagram at left in accompanying illustration). The vocal tract is terminated at one end by the lips and at the other end by the larynx, the location of the vocal cords.

During speech production, the vocal tract is excited either by pulses of air allowed to pass into the vocal tract by the vocal cords vibrating nominally in the range of 100 to 300 cycles per second, or by the noise generated from turbulent air passing through a constriction in the vocal tract (illustration, top right). The first type of excitation results in voiced speech sound such as vowels, while the second type results in unvoiced speech sound such as whispered speech and many consonants. The shape of the vocal tract is manipulated during articulation and acts like a time-varying filter, emphasizing certain frequencies and inhibiting others.

A speech synthesizer based on this model of speech production is shown in the bottom right illustration. The synthesizer consists of a time-varying filter excited by either a pulse generator or a noise source. The repetition rate of the pulse generator is controlled by a pitch period parameter. Time-varying vocal tract parameters must be supplied to control the vocal tract filter. In addition, a voicing decision has to be made to determine whether the noise or pulse source excites the vocal tract filter.

following the speech sounds. These echoes have been successfully removed by center clipping—a process that removes signals of small amplitudes while leaving large amplitude signals unaffected (see illustration, page 323, left, and Reference 1, "Additional Reading," page 325). When the echoes are of small amplitude relative to the speech waveform, they can be eliminated without losing a significant part of the speech waveform.

However, center clipping results in harmonic distortion, evident in the discontinuities of the clipped waveform. This distortion can be avoided by dividing the speech into several adjacent frequency bands using an input filter bank, center clipping each band separately, and then removing the harmonic distortion products by a set of output filters identical to the input set. The resulting speech is of good quality and long-time echoes are removed. (Listen to Bands 1 and 2 on the accompanying record.)

The reduction of reverberation is particularly striking for speech picked up in a room with long reverberation time—for example, an auditorium as demonstrated on the recording. In a small room, however, echoes of both short and long delays are present. Center clipping removes the reverberant tails, but the hollowness caused by short-delay echoes is still present after center clipping.

Removing Early Echoes

One method for removing early echoes uses two or more microphones to pick up the sound (see Reference 2, "Additional Reading"). An early echo causes some frequencies to be cancelled out in each microphone signal resulting in nulls or zeroes in the frequency response. Since the microphones are in different positions, early echoes arrive at different times in each of the microphones and consequently the resulting nulls are at different frequencies.

In this method, the microphone outputs are combined so as to minimize the effects of the frequency nulls for the particular location of the sound source. This is accomplished by dividing each microphone output into several adjacent frequency bands using a filter bank. Energy in corresponding bands is compared and the band with the greatest energy—i.e., the one least affected by nulls in the spectrum for the particular source position—is used in the combined output. This method has been shown to yield significant improvement in simple situations where only a few prominent reflections exist—for example, from a table top or wall. (Listen to Band 3 on the record.) The effects in actual rooms are still under laboratory investigation.

A different approach to the problem of rever-

THE COCKTAIL PARTY EFFECT

When several people in a room are conversing at the same time, a person can choose to concentrate on one of the talkers and hear his or her speech flow unimpeded. This remarkable ability, usually referred to as the "cocktail party" effect, results in part from binaural (two-eared) hearing. (See "Additional Reading," References 8 and 9.) In contrast, a listener with a severe hearing loss in one ear finds it difficult to attend to a particular talker under the same circumstances.

The simplest kind of telephone link between groups of people at different locations uses a microphone and loudspeaker at each location, positioned at some distance (typically a few feet) from each other and from the users. In this situation, listeners at a remote location lose the advantages of binaural listening and, like the person with a hearing loss in one ear, find it difficult to concentrate on a particular member of the conversing group. Binaural listening can be restored but only by the use of headphones at the listening end as well as two microphones and two transmission paths.

Many experiments have been carried out to help us understand and duplicate the ability to concentrate in a "cocktail party" environment. The experiments have used the outputs of two or more microphones which have been combined into a single signal by various forms of complex processing to emphasize speech coming from a particular location. Under some conditions this processing has exceeded our natural ability to discriminate a desired speech signal from a distracting background. Many of these processes, however, introduce distortion into the resulting processed signal. This distortion, combined with the difficulty of specifying in practice which signal is desired, probably makes these methods unacceptable for normal use in telecommunications. Thus, except for highly structured conference situations, the multimicrophone-simulated cocktail party effect is likely to remain a laboratory curiosity.

beration has been taken recently (see "Additional Reading," References 3 and 4). Instead of eliminating the effects of the room in the speech picked up by the microphone, this approach extracts the essential features of the speech from the reverberant signal and from it attempts to reconstruct clean (nonreverberant) speech.

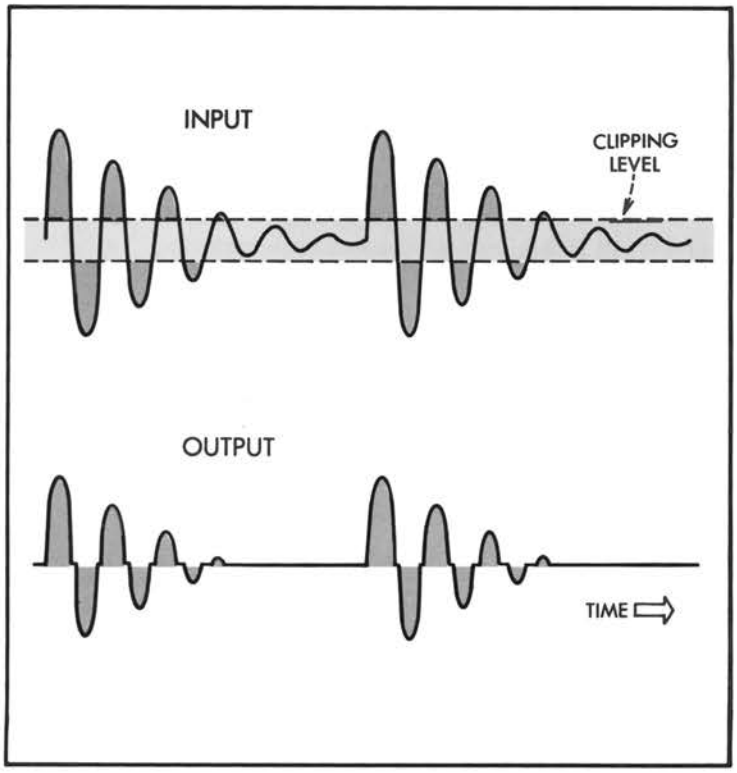
A brief review of how speech is produced in the vocal tract will help the reader to understand how this method can be successful. Speech can be characterized as "voiced" or "unvoiced"—depending on whether the source is a pulsed or noise source (see "How Speech is Produced," page 321). The parameters that determine the speech sound produced are pitch period (if the speech is voiced) and the characteristics of the vocal tract. This set of parameters can be extracted from real speech as time-varying functions. When the parameters are supplied as input to a speech synthesizer, the original speech can be reconstructed.

For speech that has been transmitted through a room, the speech parameters have to be extracted from a distorted signal. At any time, speech has a simple frequency pattern with a few broad resonances that can be described by a few parameters. On the other hand, because the room produces a great many echoes, it has a very complicated and rapidly varying frequency response (right figure, opposite page). When speech is distorted by the room, the complex room response is imposed on the relatively simple resonant structure of the speech. In the extraction of the vocal tract parameters from the reverberant speech, the effects of the room tend to average out, thus making it possible to determine vocal tract parameters which are relatively free of reverberation effects.

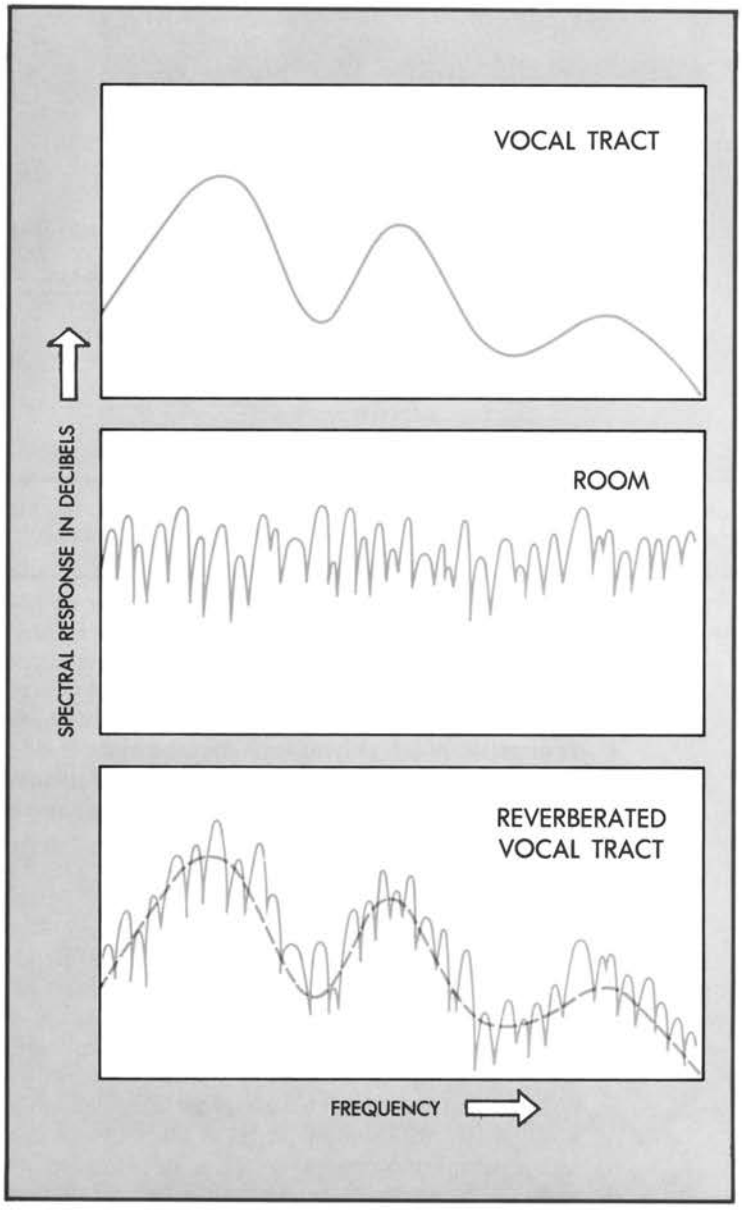
Preliminary studies have shown that good vocal tract data can in fact be extracted from reverberant speech and, under some conditions, speech with no trace of reverberation can be synthesized from the data although the speech is not yet of telephone quality because of erroneous determination of some of the speech parameters—for example, the decision whether the speech is voiced or unvoiced at a given time. (Listen to Band 4 on the record.) Work is continuing on understanding the analysis-synthesis process and on improving the difficult process of parameter extraction so as to obtain high-quality nonreverberant speech.

Now, even with the ability to transmit high-quality speech from a distant pickup, we still do not have ideal communication. With a microphone and loudspeaker in the same room, acoustic feedback and echo must be prevented without interfering with simultaneous two-way conversation.

The echo paths in such a system are illustrated



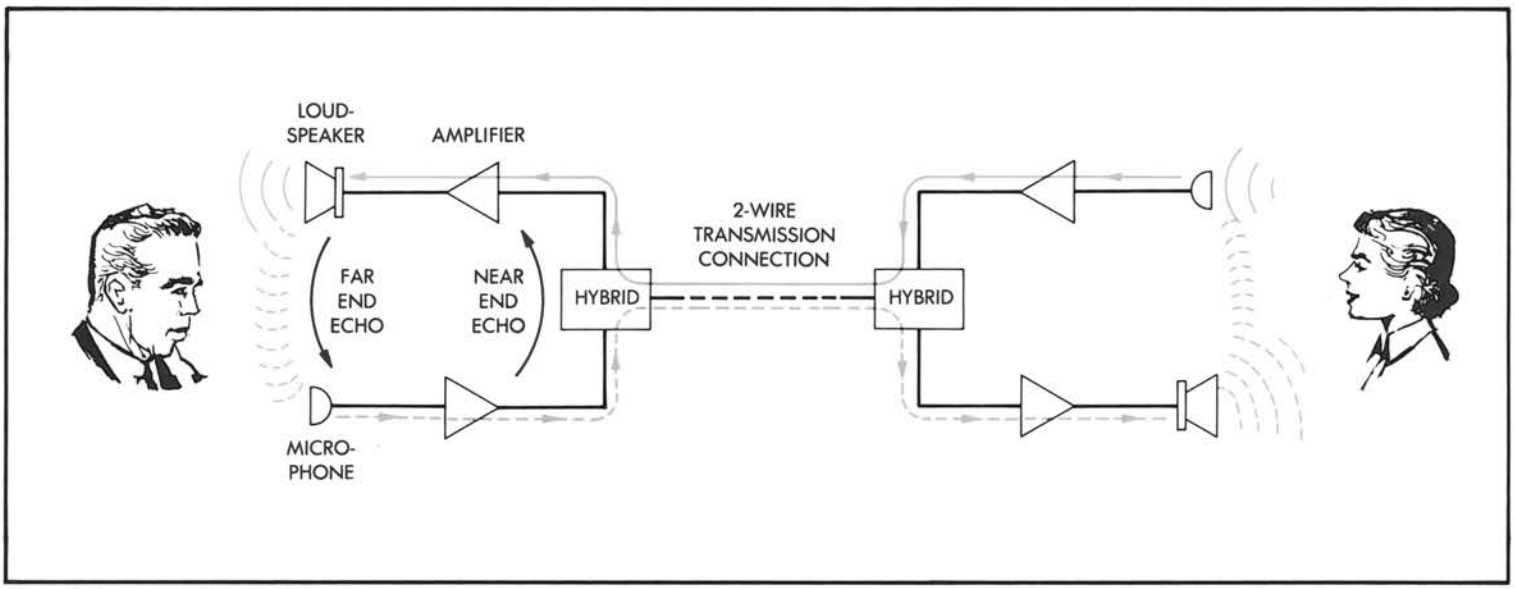
Center clipping removes signals of small amplitudes but leaves large-amplitude signals unaffected. The center-clipping technique is used to remove relatively long-delay echoes from the signal without losing a significant part of the speech waveform.



When speech is distorted by a room, the complex room response is imposed on the relatively simple resonance structure of the speech. Top: spectral response of the human vocal tract. Center: spectral response of a typical room. Bottom: the combined spectral response of the human vocal tract and a reverberant room.

on page 324. Speech from the far end is not only heard by the near-end listener but is picked up by the near-end microphone, and unless it is prevented, will return as an echo to the far-end speaker. If this echo is at a high enough level, feedback instability results. Even if the system is stable, this echo can cause a breakdown in communication when there is a considerable delay in the transmission path—as in a long-distance connection. (It is very difficult to talk in the presence of a delayed echo of one's own voice.) Another source of echo is the hybrid that connects the four-wire and two-wire section of the transmission path in the hands-free telephone. At this junction, speech from the near-end talker picked up by the near-end microphone is transmitted across the hybrid and returns to the near-end loudspeaker. The echo delay is negligible in this case, but feedback instability can result in "singing" in the short near-end loop.

Singing and echo are prevented in currently available speakerphones by use of voice switching (see "Additional Reading," Reference 5). The voice switching arrangement attenuates the received signal sufficiently to prevent singing and undesirable echo when the near-end talker is speak-



Speech from the far end (right) is heard by the listener at the near end (left), but also—unless prevented—returns to the far end as an echo. The near-end hybrid can also be a problem if it returns speech from the near-end speaker to the loudspeaker.

ing. This usually results in a situation where transmission can only effectively take place in one direction at a time. Thus, when both participants talk simultaneously ("double-talking") both speech signals cannot be transmitted; one is lost.

Echo Cancellation

A possible alternative solution to the problem of echoes would be cancellation ("Additional Reading," References 6 and 7). In this approach, a replica of the echo is generated and subtracted from the real echo in order to cancel it completely. This method of echo control permits full two-way communication in which both participants can talk and listen at the same time. It is the only method that, at high echo levels, does not in principle degrade double-talking speech.

In the case of the echo developed in the room (the far-end echo) the replica is generated by passing the signal at the loudspeaker through a network designed to duplicate the reverberation or echo properties of the room. Although this method has not been explored extensively for control of room echoes, it may be attractive for prevention of feedback and for obtaining some echo suppression. However, the long delays (greater than 200 milliseconds) required to duplicate room characteristics make it difficult to produce complete echo cancellation, and there will be a small residual echo.

To get the echo loss required for long-delay connections, the residual echo can be removed by a center-clipping process similar to that for removing room reverberation (see "Additional Reading," Reference 8). To show how this application of center clipping removes echoes, let us consider the echo of the far-end signal. In the illustration on the opposite page, two circuit components have been added at the near end of the connection: a center clipper in the microphone connection, and a circuit to control the clipping level. When a signal from the far end is detected by the control circuit, it sets the clipping level to follow the amplitude of the echo signal. The microphone signal is thus reduced to zero and the echo is removed. When there is no speech from the far end, the clipping level returns to zero so that speech from the near end is transmitted without change.

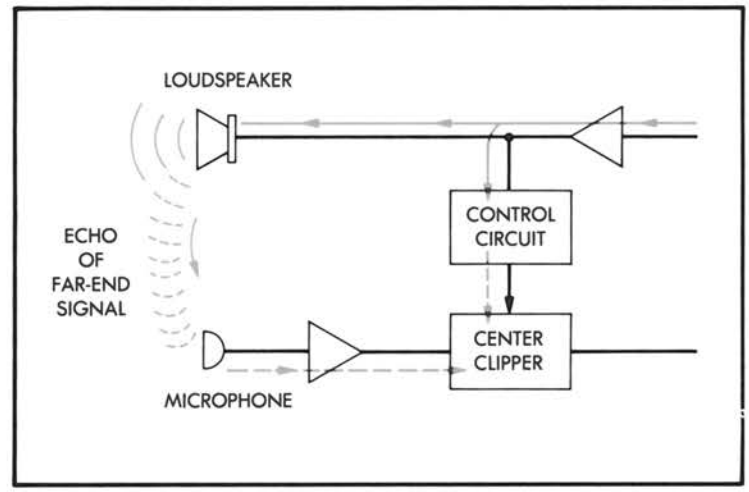
When center clipping is applied to the case of "double talking," the echo signal derived from the far-end signal is added to the signal originating at the near end. This composite signal is then center clipped. It is not intuitively obvious that center clipping will eliminate echo under these conditions. Nevertheless, it was found experimentally that no recognizable echo is ever heard by the far-end talker even when the echo is at the same level as the near-end speech. However, for high echo levels—i.e., when large clipping levels are necessary—considerable distortion of the near-end speech results from center clipping.

This discussion of echo elimination by center clipping has been somewhat simplified. To reduce distortion effects, it is advantageous to divide the

clipping up into several adjacent frequency bands as discussed earlier for removal of long-time echoes. In addition, because the loudspeaker signal is subject to room reverberation effects, the clipping levels must be maintained and allowed to return to zero at a rate comparable to the decay of echoes in the room. In the communication system shown on the opposite page, a similar center-clipping echo suppressor would also be required to remove near-end echo.

The center-clipping echo suppressor allows excellent two-way communication when the echo level in the absence of the center clipping is about 15 dB below the level of the wanted signal. Echoes of this level can be very disturbing over long distance circuits. Echo levels are usually greater, however, and must be reduced for optimum operation of the center-clipping echo suppressor. This is the function performed by the echo cancellation as previously discussed. Other ways of obtaining such presuppression are also being investigated, some based on a knowledge of the properties of speech.

At present, the types of speech processing described in this article are usually carried out in a simulated environment on a hybrid digital computer and are still too costly for practical deployment. But in the future, with increasingly ingenious and inexpensive integrated digital and analog hardware, such methods may well be within the range of our technological capability and enable us to approach the ideal in hands-free communication. □



Center clipping can remove loudspeaker-microphone coupling. The control circuit sets the clipping level to follow the amplitude of the echo in the far-end signal, thereby eliminating the echo.

ADDITIONAL READING

Long-time Reverberation

1. O. M. M. Mitchell and D. A. Berkley, "Reduction of Long-Time Reverberation by a Center-Clipping Process, *Journal of the Acoustic Society of America,* Vol. 47 (1970), p. 84 (abstract).

Short-time Reverberation

2. J. L. Flanagan and R. C. Lummis, "Signal Processing to Reduce Multipath Distortion in Small Rooms," *Journal of the Acoustic Society of America,* Vol. 47 (1970), p. 1475.

Analysis-Synthesis

3. J. B. Allen, "Synthesis of Pure Speech from a Reverberant Signal," U.S. Patent No. 3,786,188 (January 15, 1974).
4. B. S. Atal and Suzanne L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *Journal of the Acoustic Society of America,* Vol. 50 (1971), pp. 637-655.

Voice Switching

5. A. Busala, "Fundamental Considerations in the Design of a Voice-Switched Speakerphone," *Bell System Technical Journal,* Vol. 39 (1960), p. 265.

Echo Canceller

6. J. L. Kelley and B. F. Logan, "Self Adaptive Echo Canceller," U.S. Patent No. 3,500,000.
7. M. M. Sondhi, "An Adaptive Echo Canceller," *Bell System Technical Journal,* Vol. 46 (1967), pp. 497-511.

Center-clipping Echo Suppressor

8. O. M. Mracek Mitchell and David A. Berkley, "A Full-Duplex Echo Suppressor Using Center Clipping," *Bell System Technical Journal,* Vol. 50 (1971), p. 1619.

Cocktail Party Effect

9. J. F. Kaiser and E. E. David, "Reproducing the Cocktail Party Effect," *Journal of the Acoustic Society of America,* Vol. 32 (1960), p. 918 (abstract).
10. O. M. Mracek Mitchell, Carolyn A. Ross, and G. H. Yates, "Signal Processing for a Cocktail Party Effect," *Journal of the Acoustic Society of America,* Vol. 50 (1971), pp. 656-660.