BRITISH ACOUSTICAL SOCIETY: SPRING MEETING at Chelsea College
London S.W.3. on Wednesday, 25th April to Friday 27th April 1973

SPEECH AND HEARING: SESSION B: SPEECH ANALYSIS AND TRANSMISSION

Paper No.
73SHB2

SPEECH PROCESSING FOR IDEAL DISTANT-TALKING COMMUNICATION

O. M. Mracek Mitchell

Bell Laboratories, Holmdel, New Jersey 07733 U. S. A.

I. INTRODUCTION

In ideal communication between groups of people at different locations, conversation should be carried out as easily and with as high quality as if they were all in the same room. This paper discusses the problems encountered in the audio portion of such a communication system, reviews speech processing techniques which have been proposed to overcome these problems and describes new results towards achieving an ideal distant-talking telephone.

The usual configuration for a distant-talking telephone is a microphone and a loudspeaker located some distance (typically a few feet) from each other and from the users. In most cases, only a single transmission path is available for the connection between the two sets of users. This physical arrangement leads to two classes of problems, one set associated with the pickup of speech in the room by the microphone and one set associated with the coupling between the loudspeaker and the microphone.

In face-to-face communication, a binaural listener is able to concentrate on speech coming from a specific talker in a group (binaural cocktail party effect). In addition, a binaural listener has the ability to suppress the perceptual effects of room reverberation. For the remote listener in the communication situation above, where the sounds are picked up by one microphone, the ability to discriminate between different talkers is greatly reduced, and the speech from any one talker is perceptually degraded by room reverberation. Speech processing techniques to reproduce the cocktail party effect and to reduce reverberation are discussed in Sections II and III.

Inherent echo and instability exist in the distant talking telephone as a result of having both the microphone and loudspeaker in the same room. Speech processing techniques that eliminate echo and instability while allowing simultaneous 2-way communication are discussed in Section IV.

II. COCKTAIL PARTY EFFECT

Discrimination against all sound sources in a room except a selected one has been obtained by a number of previously reported techniques of processing the outputs of one microphone or of arrays of two or more microphones. Linear techniques include predictive filtering, and directional discrimination achieved by the use of directional microphones or by linear processing of microphone arrays. A nonlinear technique that has been applied by Kaiser and David[2] is cross-correlation of the outputs of two microphones in an attempt to reproduce the binaural cocktail party effect.

Another nonlinear processing scheme has been described by Mitchell, et al.[3] When this process is applied to the outputs of a microphone array, speech from a particular (on-center) location is unaffected while speech from other locations is distorted and attenuated. This processing is exactly equivalent to an instantaneous selection of one of the microphone outputs that depends on the relative ordering of the magnitudes of the outputs. However, the process described above does not minimize the contributions from unwanted sources particularly during silent intervals of the on-center source when the unwanted source may be particularly distracting. A novel process that reduces contributions of unwanted sources during pauses of the on-center source is the selection of the microphone output that is instantaneously closest to zero. For this new process, the microphone selection is not independent of the on-center source, as it was in the previously described process, but is determined by the relative magnitudes of the components of wanted and unwanted sources in each microphone. Thus, when components of both sources are present, the wanted source will suffer some degradation.

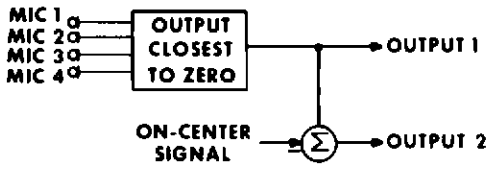

Fig. 1. Nonlinear process for microphone array

This process was computer simulated for a 4-microphone array as shown in Fig. 1. In the output of the process (Output 1), the off-center signal was completely unintelligible, even during pauses of the on-center speech. When only the on-center signal was present, the output was identical to the on-center signal. However, considerable degradation of the on-center speech resulted when energy from both sources was present at the microphones. The influence of the on-center source on the selection process is apparent at Output 2 where the on-center signal A (accessible in the simulation) has been subtracted from the processor output. The on-center signal can be plainly heard at Output 2, even though it has been subtracted from Output 1.

This process may find application in situations where it is necessary to make a particular speech signal completely unintelligible. If the signal to be rejected is of smaller amplitude as might be achieved by judicious location of the microphones, this goal might be achieved with acceptable distortion of the on-center source. However, further investigations are necessary to determine the accuracy required in location of the on-center talker and the effect of room reverberation.

III. REVERBERATION

When there is any significant spatial separation between the talker and the microphone, speech picked up by the microphone is distorted by room reverberation. The degradation of the speech can be separated into two perceptual effects. One of these is the spectral distortion called "coloration" which results from room echoes that are separated by relatively short delay times. The second is time distortion due to temporally resolved echoes at relatively long delay times.

In the first effect, the spectrum is characterized by sharp peaks and valleys caused by superposition of echoes with small delay differences. Flanagan and Lummis[4] have described a technique for combining the signals from two or more microphones to produce a flatter spectrum. The second effect has been reduced by application of center clipping in contiguous narrow frequency bands.[5] The center clipping removes reverberant tails of speech caused by echoes at long delay times. These techniques are effective under limited conditions, the former when there is a prominent zero due to an early room reflection and the latter in large rooms where the echoes are well separated and reduced in amplitude.

A recent breakthrough in dereverberation has been made by Allen.[6] While previous techniques attempted to remove the room effects from the microphone signal, this technique attempts to extract clean speech from the degraded signal. The success of this method is based on the fact that the transfer function of a room has a generally complicated structure while speech can be described by relatively few parameters. Allen uses an analysis-synthesis scheme based on a linear prediction vocoder similar to that described by Atal[7] as shown in the simplified block diagram of Fig. 2. The parameters extracted are voiced-unvoiced decision, pitch period during voicing, amplitude of excitation function and the poles of an all-pole vocal tract model.

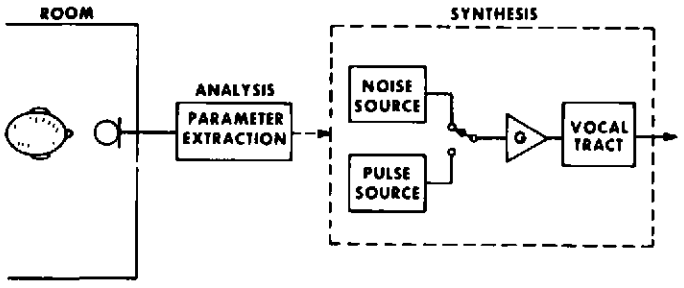

Fig. 2. Linear prediction vocoder for dereverberation (after Allen[6])

Preliminary experiments show that the parameters can be extracted from reverberant speech signals and that natural-sounding dereverberant speech can be generated by the synthesizer. Application of this process to the distant-talking telephone is discussed at the end of the next section.

IV. ECHO AND INSTABILITY

A. Sources of Echo and Instability

In an ideal distant-talking telephone, echo and instability must be prevented by a technique that allows simultaneous 2-way conversation. High-quality 2-way conversation should be possible even when there is a satellite transmission delay between the ends of the connection.

Fig. 3 is a schematic diagram of two distant-talking telephones connected by a 2-wire transmission path that shows the two types of feedback paths that exist. One of these is the acoustical coupling between the loudspeaker and the microphone which causes an echo of the far-end signal to return to the far-end loudspeaker. The second is the electrical path through the hybrid H which causes an echo of the near-end signal to return to the near-end loudspeaker. Two possible types of feedback loops then exist: first, an end-to end loop through the loudspeaker-microphone

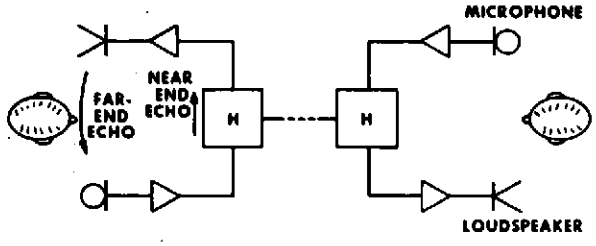

Fig. 3. Distant-talking telephones connected by a 2-wire transmission path showing echo paths

coupling at each end of the connection and, second, a near-end loop including the microphone-loudspeaker coupling and the hybrid path at one end of the connection. For typical listening and speaking levels the gain in each of these feedback loops is in the vicinity of 0 dB, so that high echo levels and instability exist.

B. Center Clipping Echo Suppressor

One method currently known to be effective for providing stability and echo suppression while allowing full-duplex (simultaneous 2-way) operation is the center-clipping echo suppressor.[8] This system provides stability even in the presence of loop gain. However, it is full-duplex only when the echo levels are sufficiently low.

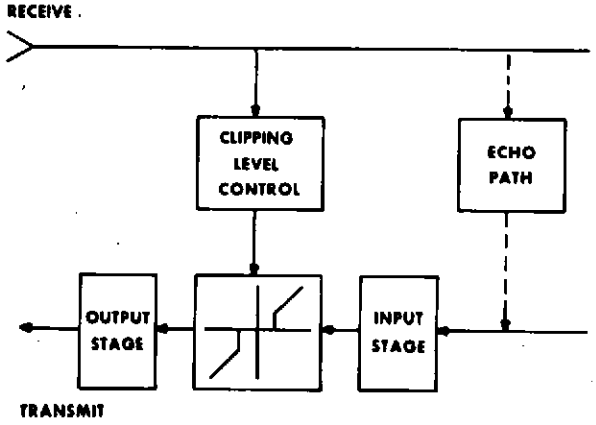

Fig. 4. Center clipping echo suppressor

Figure 4 is a simplified block diagram of the center-clipping echo suppressor in a circuit where an echo path, either acoustical or electrical, exists between the receive and transmit paths. The center-clipping stage in the transmit path removes the echo signal when the clipping level is controlled by the signal in the receive path. In a multiband center-clipping echo suppressor,[8] both input and output stages are banks of contiguous band-pass filters which prevent harmonic distortion products produced by center-clipping from appearing at the output. The center-clipping stage consists of a set of center clippers with independent clipping level control, one in each frequency band. With transmission delay of 600 msec and hold-over times of 10 msec, this echo suppressor is almost indistinguishable from a full-duplex connection when the echo level is 15 dB below the transmitted signal level.

C. Presuppression

Since the center-clipping echo suppressor works well only in the presence of significant return loss, some form of full-duplex preprocessing has to be used to reduce the echo signal by at least 10 dB. The techniques described in Section II are applicable to this problem since the loudspeaker is another source to be discriminated against.

Additional techniques can be applied in this case as a consequence of having the loudspeaker signal available for measurement or modification before it is transmitted to the microphone through the room acoustics. Thus, comb filtering is one technique that can be used to reduce the coupling.[9] In this arrangement, a comb filter is introduced into the loudspeaker path while the complementary comb filter is introduced into the microphone path. A time-domain analogue to this process in which the loudspeaker and microphone signals are alternately switched at an ultrasonic rate has been proposed.[10]

Another possibility for presuppression is canceling the echo signal[11] at the microphone by constructing a replica of the microphone signal from the signal at the loudspeaker. The replica is generated by passing the loudspeaker signal through a filter that has been adapted to match the transfer function of the room.[12]

The predictive filtering technique that has been described previously[1] can be used for presuppression. Suppression of speech during voiced segments is achieved by setting a predictive filter to reject the pitch frequency and its harmonics. Results of simulations using this form of presuppression are discussed below.

D. Predictive Filtering and Center Clipping for Echo Suppression

Predictive filtering presuppression has been investigated in simulation in conjunction with a center-clipping echo suppressor.[13] Figure 5 shows the configuration simulated for removing far-end echo in a distant talking telephone. Pitch information extracted from the incoming far-end signal is used to set the filter in the return path to reject the unwanted speech coupled from the loudspeaker to the microphone. An identical predictive filter reduces the clipping levels required to remove the residual loudspeaker signal from the transmitted signal.

In the simulation, equal levels of far-end and near-end signal were used at the microphone. With this simulated return loss of 0 dB, noticeable improvement in the

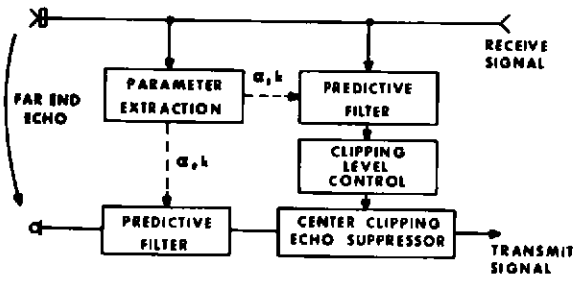

Fig. 5. Suppression of far-end echo by predictive filtering and center-clipping

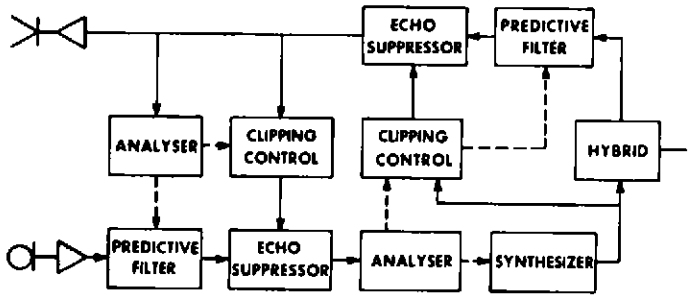

Fig. 6. A possible ideal distant-talking telephone

quality of the transmitted signal resulted from the predictive filtering when the clipping levels were controlled by the predictively filtered far-end signal. This advantage is reduced when the clipping levels are held over, as would be necessary under reverberant conditions.

A possible combination of the analysis-synthesis dereverberation scheme with this form of echo suppression for both near-end and far-end echo is shown in Fig. 6. The arrangement for the far-end echo is the same as in Fig. 5. Since the near-end echo is derived from the synthesized (dereverberated) speech, the parameters for the predictive filter in the near-end echo suppressor are supplied by the analyzer in the vocoder.

V. CONCLUSIONS

The problems present in a distant-talking communication system have been outlined and a number of speech processing solutions have been discussed. At present this processing has been limited to experimental situations and simulations, and outstanding improvements have been produced. These results demonstrate the possibility of achieving much higher quality distant-talking communication.

ACKNOWLEDGMENTS

I wish to thank Dr. D. A. Berkley for his interest and participation in this work and Dr. J. S. Courtney-Pratt for his continued support.

VI. REFERENCES

1. O. M. M. Mitchell, "Signal Processing for Multimicrophone and Single Microphone Cocktail Party Effects," Proceedings of the Seventh International Congress on Acoustics, (Budapest 1971), Paper 24C10.
2. J. F. Kaiser and E. E. David, Jr., "Reproducing the Cocktail Party Effect," J. Acoust. Soc. Amer. 32, (1960), p. 918(A).
3. O. M. M. Mitchell, C. A. Ross and G. H. Yates, "Signal Processing for a Cocktail Party Effect," J. Acoust. Soc. Amer. 50, No. 2 (Part 2), (August 1971), pp. 656-660.
4. J. L. Flanagan and R. C. Lummis, "Signal Processing to Reduce Multipath Distortion in Small Rooms," J. Acoust. Soc. Amer. 47, No. 6 (Part 1), (1970), pp. 1475-1481.
5. O. M. M. Mitchell and D. A. Berkley, "Reduction of Long-Time Reverberation by a Center-Clipping Process," J. Acoust. Soc. Amer. 47, No. 1 (Part 1), (1970) p. 84.
6. J. B. Allen, "Speech Dereverberation," Eighty-Fourth Meeting of the Acoustical Society of America, Paper Q13 (November 1972).
7. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer. 50, (1971), pp. 637-655.
8. O. M. M. Mitchell and D. A. Berkley, "A Full-Duplex Echo Suppressor Using Center Clipping," B.S.T.J. 50, No. 5 (May-June 1971), pp. 1619-1630.
9. J. S. Courtney-Pratt, private communication; D. A. Berkley and J. S. Courtney-Pratt, "Telephone Transmission Using Complementary Comb Filters," U. S. Patent No. 3,622,714 (November 23, 1971).
10. D. Mitchell, "Switching Circuit for Cancelling the Direct Sound Transmission from the Loudspeaker to the Microphone in a Loudspeaking Telephone Set," U. S. Patent No. 3,601,549.
11. M. M. Sondhi, "An Adaptive Echo Canceller," B.S.T.J. 46, (Mar 1967), pp.497-511.
12. D. A. Berkley, private communication.
13. G. H. Yates, private communication.