Signal Processing for a Cocktail Party Effect

O. M. Mracek Mitchell, Carolyn A. Ross, and G. H. Yates

Citation: The Journal of the Acoustical Society of America **50**, 656 (1971);

View online: https://doi.org/10.1121/1.1912680

View Table of Contents: http://asa.scitation.org/toc/jas/50/2B

Published by the Acoustical Society of America

9.8, 9.5

Received 2 October 1970; revised 14 May 1971

Signal Processing for a Cocktail Party Effect

O. M. Mracek Mitchell, Carolyn A. Ross, and G. H. Yates

Bell Telephone Laboratories, Holmdel, New Jersey 07733

A binaural listener has the ability to concentrate on speech from a particular location while suppressing speech from other locations (binaural "cocktail party" effect). In some communication situations where sounds are picked up by a microphone system for transmission on a single path to a remote listener, it would be desirable to preprocess the signals to achieve a similar effect. We describe a class of nonlinear processes for the outputs of an array of microphones which emphasize speech coming from a particular (on-center) location in a background of other sounds. These processes completely eliminate any off-center impulsive noise which is nonoverlapping at the four microphones. Results of processing outputs of real and computer-simulated microphone arrays for speech and noise signals are described. Under anechoic conditions, the processing results in reproduction of the on-center speech without change, and in distortion and attenuation of an off-center speech source. The distortion produced by the processing appears to be an important factor in subjective suppression of the off-center source.

INTRODUCTION

A listener present in a room with several sources of sound is able to concentrate on any one of the sound sources. This emphasis of a particular sound source and rejection of the other sound sources has been referred to as the binaural "cocktail party" effect. In many communication situations, sound must be picked up by a microphone system for transmission on a single path to a remote listener. If two or more sound sources are within the range of a single microphone, their signals are combined in the microphone output and, if transmitted directly, appear jumbled to the remote listener; that is, he no longer has the advantage of the binaural cocktail party effect. In such a situation, it would be desirable to preprocess the acoustical signals so as to emphasize the speech of a particular person in a background of other sounds.

Previously, Kaiser and David[1] have reproduced the cocktail party effect by applying nonlinear cross-correlation techniques to the outputs of two microphones. In this paper, we discuss in detail a class of nonlinear processes using a microphone array which emphasizes a wanted signal relative to unwanted signals from other locations. The unwanted signals (either speech or certain kinds of random noise) are attenuated and distorted, while the wanted signal is unaffected. When the unwanted signal is speech, the distortion makes it less intelligible.

I. NONLINEAR PROCESSING

The block diagram of such a nonlinear process for four microphone outputs is shown in Fig. 1. We first discuss the operation of this system for impulsive sound sources. Consider a wanted sound source S located at the center of the microphone array and an unwanted sound source N at an off-center location. Each source emits short bursts of sound which are picked up by the four microphones. The sound from S reaches all four microphones simultaneously, producing an identical signal S_1 in each microphone. For the case shown, the sound from N reaches the four microphones at different times, producing signals N_1, N_2, N_3, and N_4 at the four microphones. For simplicity, the change of amplitude with distance is neglected. The waveforms at microphones 1 and 2 are shown schematically.

The outputs of the microphones taken in pairs are processed by two identical first stages consisting of three operations (subtract, positive full-wave rectify, and add). Waveforms are shown after each operation in the process for one of the first stages. The first step in the process takes the difference between the two microphone outputs, S_1+N_1 and S_1+N_2, thus cancelling the on-center signal S_1. The difference signal N_1-N_2, containing only components of the off-center source N, is then full-wave rectified and added to the sum of the two microphone outputs. When this signal is attenuated by 6 dB, the output is $S_1+\frac{1}{2}(N_1+N_2+|N_1-N_2|)$. For the special case shown in Fig. 1 where signals N_1 and N_2

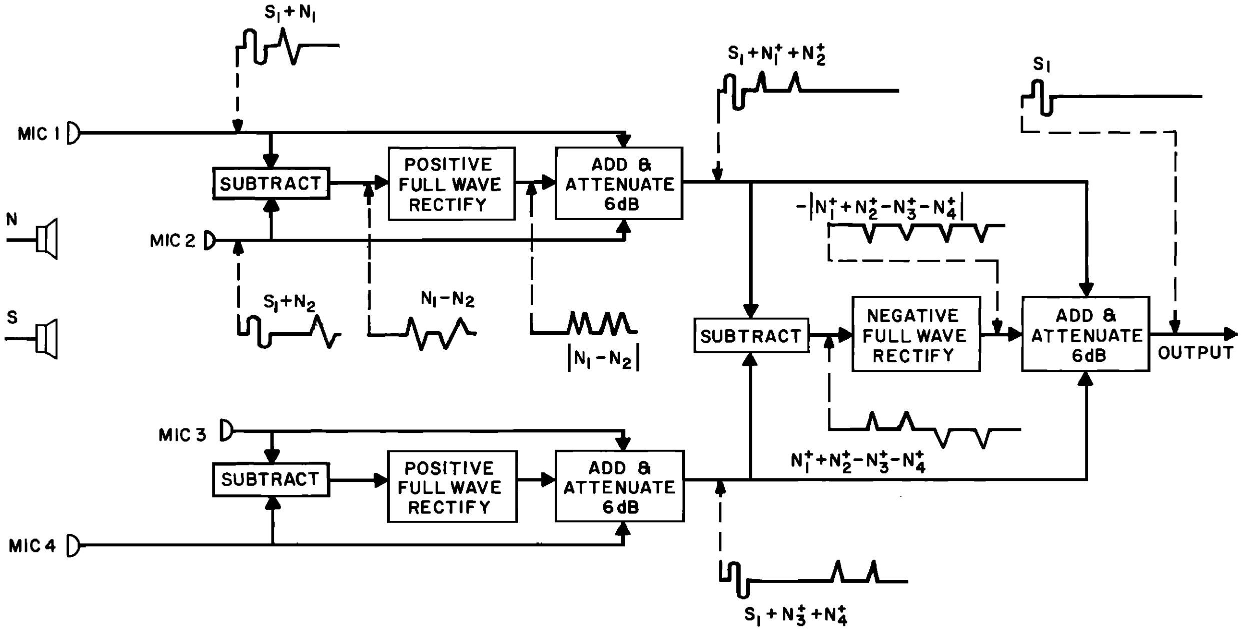

FIG. 1. Block diagram of a two-stage nonlinear signal-processing system. Schematic (nonoverlapping) signals are shown at various points of the system.

do not overlap in time, this expression simplifies to $S_1+N_1^++N_2^+$. (N^+ is the positive part of signal N.) Thus, for nonoverlapping signals, the result of processing by a stage with a positive full-wave rectifier is to cancel negative-going parts of the off-center source. The resulting waveform consists of an undistorted wanted signal and positive portions of the unwanted signals. A similar output $S_1+N_3^++N_4^+$ is obtained by applying this process to the outputs of microphones 3 and 4.

The outputs of the two first stages are then processed by a second stage which differs from the first stages only in reversal of the sign of the full-wave rectifier output. The result of processing nonoverlapping signals by a stage with a negative full-wave rectifier is to cancel positive-going portions of the off-center source. For the case shown in which N_1, N_2, N_3, and N_4 do not overlap in time, the inputs to the second stage contain only positive-going portions of the off-center source, and therefore the off-center source is completely cancelled at the output. Thus, impulsive noise which does not overlap at the four microphones is eliminated by this nonlinear process.

When the off-center signal is speech, part of the signal will overlap at the microphones. This part of the unwanted signal will not be eliminated, but will be distorted by the rectifying stages of the processor. The resulting attenuation and distortion of unwanted speech is discussed in the next section.

The output of this processor appears to be a complicated function of all four inputs. It has been pointed out, however, that the instantaneous processor output is always exactly equal to *one* of its inputs.[2] The two-stage processor of Fig. 1, in effect, looks at only one input at a time and ignores all the others.

This result can be seen by inspecting the output of any one of the stages in the process. If the inputs to the first stage are E_1 and E_2, the output is $\frac{1}{2}(E_1+E_2+|E_1-E_2|)$. This expression is equal to the greater of E_1 and E_2; that is, the output of a stage with a positive full-wave rectifier is the greater of the two inputs. Similarly, if the inputs to the second stage are E_3 and E_4, the output is $\frac{1}{2}(E_3+E_4-|E_3-E_4|)$. This expression is equal to the lesser of E_3 and E_4; that is, the output of a stage with a negative full-wave rectifier is the lesser of its two inputs.

From these considerations, it can be seen that the block diagram of Fig. 1 is functionally equivalent to the simplified block diagram of Fig. 2(a), where the two first stages are replaced by maximum operations and the second stage is replaced by a minimum operation. This process is referred to as the MAXMIN process. It can be seen that the MAXMIN process eliminates the instantaneous (algebraic) maximum and minimum microphone outputs. Thus, the processor output is one of the remaining two microphone outputs, either the second or third largest.

The equivalence of the circuit of Fig. 1 to the circuit of Fig. 2(a) allows an alternative approach to understanding the effect of the process on impulsive noise. Impulsive off-center noise of short enough duration will be received by only one microphone at a time, and will cause that microphone to have either the maximum or minimum output at that instant. Since the processor rejects the maximum and minimum microphone outputs, nonoverlapping noise is eliminated.

The circuit shown in Fig. 2(b), which is referred to as the MINMAX process, also eliminates the instantaneous maximum and minimum microphone outputs and

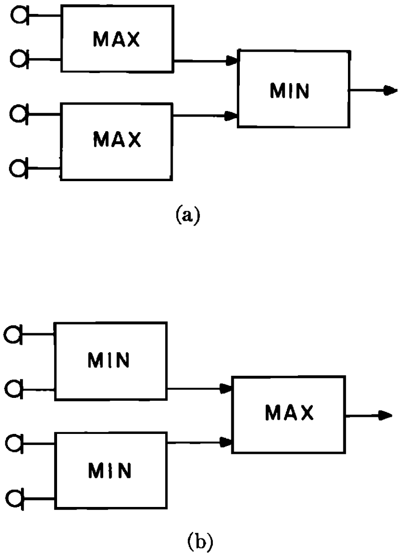

FIG. 2. (a) Block diagram functionally equivalent to the block diagram of Fig. 1 and referred to as the MAXMIN process. (b) An alternative block diagram equivalent in effectiveness to block diagram of (a) and referred to as the MINMAX process.

selects either the second or third largest microphone output. It is easily shown that, when the MAXMIN processor selects the second largest microphone output, the MINMAX processor selects the third largest and vice versa. If these two signals are independent, increased suppression will result from averaging them. The average of the middle half of a distribution has been called the "midmean" by Tukey.[3] The MIDMEAN processor also has the property of rejecting impulsive off-center sounds of either sign.

Wallace[4] has pointed out that these three nonlinear processes can be regarded as members of a general class of processors which eliminate impulsive noise of short enough duration. Two other processes belonging to this class are discussed in this paper. One of these always selects the second largest microphone output and the other always selects the third largest microphone output.

In the next section, we describe the results of processing the outputs of arrays of four microphones by the various nonlinear processes discussed in this section.

II. RESULTS

The effectiveness of this kind of signal processing in providing discrimination against unwanted sound sources can best be evaluated by subjective listening tests. Quantitative evaluation has as yet not been carried out for simultaneous on- and off-center sources. Previously, we have played examples of the nonlinear processing,[5] and we include some of these on the accompanying recording. We also present quantitative data for the amplitude suppression resulting from nonlinear processing for the case of a single off-center sound source.

We first describe results obtained with a real microphone array for two of the nonlinear processes (MAXMIN and MIDMEAN). Then we discuss results obtained by digital simulation for four of the nonlinear processes described in Sec. I. In each case, we compare the results of the nonlinear processing to those obtained by linear superposition of the microphone outputs.

A. Analog Processing of a Real Microphone Array

We investigated the MAXMIN and MIDMEAN processors using an array of four microphones and two speech sources in an anechoic chamber. This experiment shows the basic features of the processing which may hold under less ideal conditions, where wall echoes contribute to the microphone signals.

Figure 3 is a floor plan of the experimental arrangement used. Four omnidirectional microphones (Altec Lansing model 633A) indicated by circles were located at the corners of a 10-ft × 12-ft rectangle and 5 ft above the floor. The speech sources were prerecorded tapes of text read by a female voice and a male voice played at about equal volume through loudspeakers (KLH, model 12-5) located on the floor and directed at the ceiling, i.e., with symmetrical patterns in the horizontal plane. One loudspeaker was fixed at the center of the rectangle (filled-in triangle), and one was moved to various off-center positions, as shown in the diagram by the open triangles. In addition to processing by two nonlinear processors, the four microphones were also processed by linear superposition. The processing was effected by circuits constructed from commercial operational amplifiers. Experiments were done with simultaneous on- and off-center sound sources, and with a single source in various positions.

Representative examples of this processing for simultaneous talking are given on the accompanying recording. Table I gives the contents of the five bands. For bands 1–4, the female speech source was located on-center and the male source off-center at the point indicated by the arrow in Fig. 3. For band 5, the speech sources were interchanged.

In the unprocessed output of one of the microphones (band 1), the two texts are jumbled together and it is difficult to follow either speaker. Linear superposition of the microphone outputs reduces the amplitude of the off-center speech, but does not introduce any distortion of this unwanted speech (band 2). In addition, for some locations of the off-center source near any of the microphones, the amplitude of the off-center speech actually increases. In contrast, processing by either the MAXMIN or MIDMEAN processor (bands 3–5) results in both attenuation and distortion of the unwanted speech. The distortion of the unwanted speech appears to increase the intelligibility of the wanted speech. The distortion of the off-center speech is somewhat greater

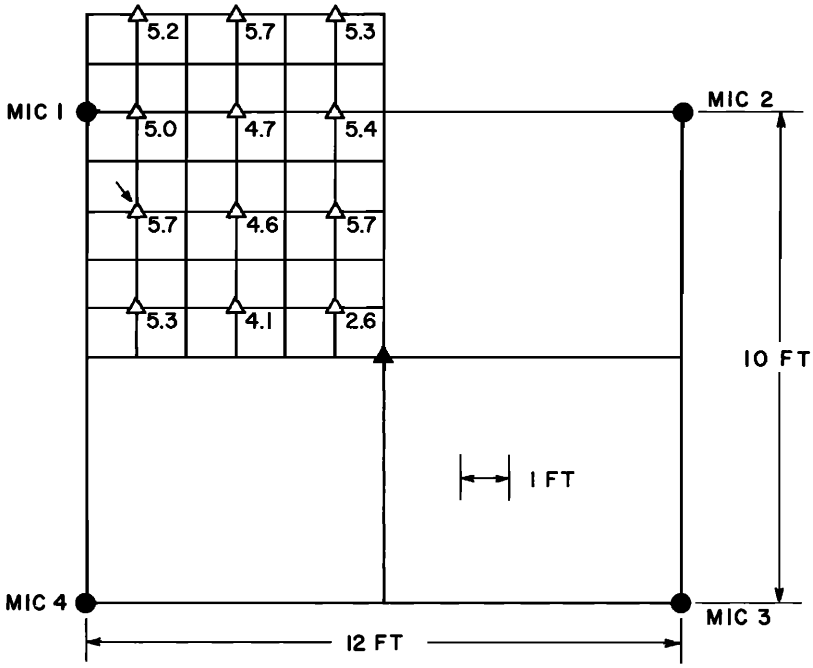

FIG. 3. Floor plan of experimental arrangement showing positions of the four microphones (●) and the positions of the on-center (▲) and off-center (△) speakers. The number shown near each off-center speaker location is the suppression in decibels obtained with the MAXMIN processor for that off-center location relative to the on-center location. Delays corresponding to the off-center location indicated by the arrow were used in the computer simulation described in Sec. II-B.

for the MAXMIN process than for the MIDMEAN, but the MIDMEAN process reduces the amplitude of the off-center speech by a larger amount.

The amount of amplitude suppression resulting from the MAXMIN process was measured for both off-center speech and impulsive noise sources. Figure 3 shows the suppression for various off-center locations of a speech source. Suppressions in the range 2.6–5.7 dB were obtained, where the smallest suppression corresponds to a location near the center of the array. For impulsive noise of about 1-msec duration from the off-center loudspeaker, amplitude suppression up to 20 dB resulted from the nonlinear processing.

B. Computer-Simulated Microphone Array

A digital computer simulation was carried out to allow a comparison of the amplitude suppression resulting from each of the nonlinear processes described in Sec. I. The outputs of the four-microphone array of Fig. 3 were simulated from digitized speech signals and from computer-generated white noise using delays corresponding to the off-center source location marked by an arrow. In this simulation, no correction to the amplitude was made for the distance between the sound source and the microphones. The rms amplitude of the output of each process was computed for the sound source located at both the on-center and off-center positions. The suppression of the off-center source relative to the on-center source was calculated in decibels from these two rms amplitudes.

Table II shows the suppression given by the various processes for white-noise and speech sources. For white noise, suppressions in the range 3.4–6.0 dB were obtained. The largest suppression of 6.0 dB resulted from the linear superposition of the microphone outputs. This is the suppression expected for linear processing of four uncorrelated signals. However, as mentioned previously, when distance effects are included, the linear process gives less suppression, and for some positions an enhancement of the off-center source. The processes that select as output one of the microphone signals resulted in about equal amounts of

TABLE I. Contents of accompanying recording.

Band	Process	
1	One microphone	Female speech source on center, male off center.
2	Linear	
3	MINMAX	
4	MIDMEAN	
5	MIDMEAN	Sources interchanged.

TABLE II. Suppression of an off-center sound source relative to an on-center source obtained in a digital simulation of a four-microphone array.

Sound source	Process	Suppression (dB)
White noise	MAXMIN	3.5
	Second largest	3.6
	Third largest	3.4
	MIDMEAN	5.2
	Linear	6.0
Speech	MAXMIN	3.7
	Second largest	3.8
	Third largest	4.0
	MIDMEAN	5.7
	Linear	6.3

suppression (3.4–3.6 dB) of the off-center white-noise source. This amount of suppression agrees with that calculated by Wallace for the MAXMIN processor with uncorrelated Gaussian inputs.[4] The MIDMEAN process, which is an average of two of the microphone outputs, gives nearly as much suppression (5.2 dB) as the linear process when no account is taken of distance effects, and more when a square-law distance correction is made. For the off-center speech source, slightly higher suppression was obtained for each process compared to the corresponding suppression of the white-noise source. The additional suppression of the nonlinear processes is probably due to the "peaky" character of speech, which results in nonoverlapping signals at the microphones and consequent elimination. The additional suppression of the linear process is probably due to correlations between the microphones.

III. DISCUSSION

The experiments described in the preceding section are idealized situations for several reasons. One of these is the fact that the effect of room reverberation was not included. In situations where room reverberation is not negligible, the reverberant components will be treated by the processor as an off-center sound source, and hence will add to the unwanted background.

Another idealization was the use of stationary loudspeakers as sound sources, which facilitated the adjustment of the delays and gains in the various microphone channels to make the signal from the on-center source identical at all the microphones. It is obviously very important for the signal from the on-center source to be identical at all of the inputs of the signal processor so that complete cancellation of the wanted signal results from the first subtraction in each stage. If the signals are not identical, part of the on-center source will be distorted by the nonlinear process. In practice, with moving sources, it may be possible to use an ultrasonic beacon[6] carried by the wanted speaker to aid in adjusting the delays and gains in the microphone channels. Cross-correlation techniques could then be used in making the final adjustments.

IV. CONCLUSIONS

We have investigated nonlinear processing, which can be used in conjunction with an array of microphones to enhance the intelligibility of speech coming from a particular location. Some of its properties are summarized here. Under nonreverberant conditions, speech from an on-center location is reproduced without change, while speech from other locations is distorted and reduced in amplitude. Impulsive sounds from other locations are eliminated from the output when they do not overlap in time at the microphones. In contrast, linear processing gives slightly better amplitude suppression of the unwanted speaker than the nonlinear processing when distance effects are negligible, but produces no distortion of the unwanted speaker. When the unwanted source is near one of the microphones, linear processing is much less effective than the nonlinear processing.

ACKNOWLEDGMENTS

We wish to thank: E. Eisner and J. S. Courtney-Pratt for suggesting use of microphone arrays for this application; M. R. Schroeder, R. L. Wallace, R. N. Thurston, and R. O'Regan for helpful discussions; D. A. Berkeley for critical comments on the manuscript; F. K. Harvey and J. G. Cisek for their loan of most of the equipment used in the early part of this investigation; J. J. Dondero for use of the anechoic chamber; and S. L. Hanauer for advice on the computer simulations.

[1] J. F. Kaiser and E. E. David, Jr., "Reproducing the Cocktail Party Effect," J. Acoust. Soc. Amer. **32**, 918(A) (1960).

[2] M. M. Sondhi and R. L. Wallace have independently pointed out this property of the nonlinear process.

[3] J. W. Tukey, "Some Perspectives on Data Analysis," J. Amer. Statist. Ass., R. A. Fisher Mem. Lecture, Amer. Statist. Ass., Washington, D. C. (Dec. 1967).

[4] R. L. Wallace (unpublished results).

[5] O. M. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal Processing for a Cocktail Party Effect," J. Acoust. Soc. Amer. **45**, 315(A) (1969).

[6] J. S. Courtney-Pratt (private communication).