Multimicrophone signal-processing technique to remove room reverberation from speech signals

J. B. Allen, D. A. Berkley, and J. Blauert[a)]

Bell Laboratories, Murray Hill, New Jersey 07974

(Received 7 March 1977)

It is well known that room reverberation can significantly impair one's perception of sounds recorded by a microphone in that room. Acoustic recordings produced in untreated rooms are characterized by a hollow echolike quality resulting from not locating the microphone close to the source. In this paper we discuss a multimicrophone digital processing scheme for removing much of the degrading distortion. To accomplish this the individual microphone signals are divided into frequency bands whose corresponding outputs are cophased (delay differences are compensated) and added. Then the gain of each resulting band is set based on the cross correlation between corresponding microphone signals in that band. The reconstructed broadband speech is perceived with considerably reduced reverberation.

PACS numbers: 43.55.Br, 43.55.Jz, 43.70.Lw, 43.60.Gk

INTRODUCTION

It is well known that room reverberation can significantly reduce the perceived quality and in some cases reduce the intelligibility of speech recorded by a microphone in the room.[1-3] This phenomenon can be important in conference telephony where the nature of the rooms used is not generally well controlled. Considerable work has been done in an attempt to improve the resulting transmission quality.[4,5] However, no single scheme has been demonstrated that is effective over a wide range of reverberation times, noise conditions, and speakers.

In general, perception of reverberation has been heuristically separated into two (or more) categories. Early room echoes are perceived as spectral distortion and their effect is known as coloration. Longer term reverberation contributes time-domain noiselike perceptions or tails on speech signals.

Reverberation-reduction processes may generally be divided into single or multiple microphone methods and into those primarily affecting coloration or those affecting reverberant tails.

In this paper we describe a multiple-microphone process which is effective for both reverberant degradations. Coloration is reduced by a banded processor generically related to that of Flanagan.[4] Long term reverberation is reduced by controlling the gain of individual bands based on the correlation between channels. This takes advantage of the uncorrelated nature of reverberant tails found previously.[6,7]

It is not known whether coloration or long-term echo is more important perceptually, and it is undoubtably true that their relative importance depends on the particular room conditions and on the source and receiver locations. In any case our present theoretical understanding of these questions is minimal and conjectural.

The overall processor was simulated on a digital computer and tested using speech recorded in a variety of actual rooms.

I. PROCESS DESCRIPTION

Assume we have signals defined by the situation shown in Fig. 1 where $s(t)$ is some undistorted source which might be music or speech in a room. Two reverberant microphone signals $x(t)$ and $y(t)$ which are mathematically given by

$$x(t)=\int_{-\infty}^{t} h(t-\tau)s(\tau)d\tau, \tag{1}$$

$$y(t)=\int_{-\infty}^{t} g(t-\tau)s(\tau)d\tau, \tag{2}$$

where $h(t)$ and $g(t)$ are the room-impulse-response functions. These functions are assumed to be unknown; however, some generalizations may be made. We know from measurements that the "tails" of g and h are uncorrelated, namely they are much like two independent noise signals.[6] We also know that the first few echoes do not arrive at the two microphone positions simultaneously.

As previously mentioned the processor works in frequency bands. In each band it first removes delays that exist between the coherent part or early echoes by a phase shift and adds the phase corrected signals. This part of the process is referred to as "cophase and add in bands." It then adjusts the gain for each band by a normalized cross-correlation function. This gain switching has the effect of turning off uncorrelated signals and passing correlated signals.

The resulting signals in each band are then combined to synthesize the signal estimate $\hat{s}(t)$. The process is outlined in block form in Fig. 2.

Before discussing the details of the various elements of the process it is worth understanding the response of the processor to various inputs. When x and y are iden-

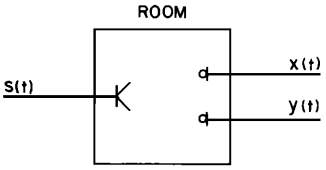

FIG. 1. Basic room configuration.

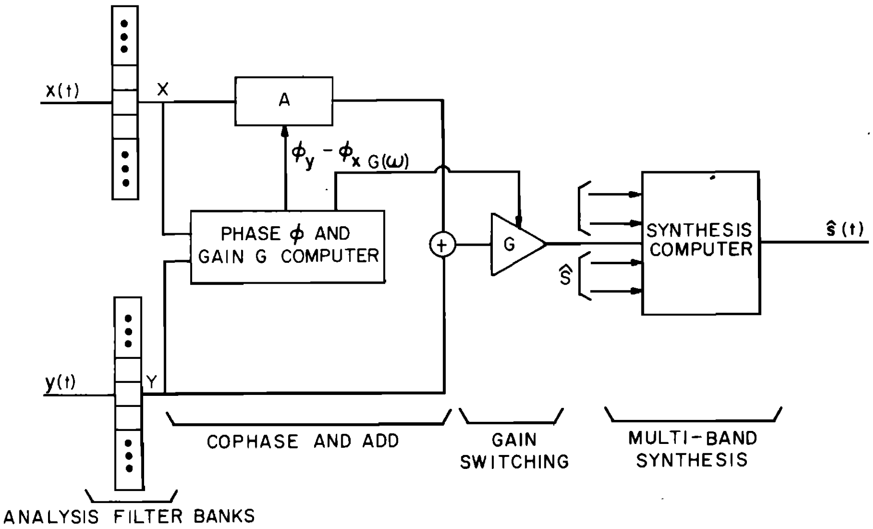

FIG. 2. Reverberation processor block diagram.

tical, $\hat{s}=x=y$. If either x or y is delayed by less than a few milliseconds the delay is corrected automatically and $\hat{s}(t)=x(t-T)=y(t)$, where T is a hypothetical delay correction. When the delay is greater than the window length L (to be defined) the x and y signals begins to appear uncorrelated and the gain tends toward zero. Finally, if x and y are totally uncorrelated, $\hat{s}(t)$ is zero. All of the above is true in each frequency band. Thus if x and y are uncorrelated at one frequency and correlated at another, then the uncorrelated part is removed and the correlated part remains. Since all of the operations are defined on a moving average basis, everything may change as a function of time.

II. PROCESSOR IMPLEMENTATION

The heart of the processor in Fig. 2 is based on a sampled data filter bank analysis/synthesis technique allowing time varying modifications to be made to the short-term spectrum.[8,9] A quick review of the analysis/synthesis formulas is in order.

We define a window function $w(n)$ which is a low-pass filter sequence such as a Hamming window. We define the sample period of the sequence as D. For speech, we have used $D=0.0001$ sec corresponding to a 5 kHz speech bandwidth. Since $w(n)$ is a low-pass sequence, any sequence $x(n)$ filtered by $w(n)$ may be resampled at a lower rate equal to twice the highest important frequency of $w(n)$. We call this new sample period T. For a Hamming window, a reasonable definition for T turns out to be a quarter of the window length. Thus if the window is L points long,

$$T=\tfrac{1}{4}LD \text{ sec.} \tag{3}$$

The notation can become confusing because we are talking about sequences being sampled at several sample rates. For this reason we will not suppress the sample period in our notation. Namely $x(n)$ will be explicitly written as $x(nD)$. A function sampled at the lower rate T would be $x(kT)$, where k is an integer and T is greater than D.

The analysis rule as discussed in Ref. 8 may be written

$$X(mF,\ kT)=\mathfrak{F}[w(nD-kT)\,x(nD)], \tag{4}$$

where

$$w(t)=\begin{cases}0.54+0.46\cos[\pi(2t-D)/(L-1)D], & D(1-\tfrac{1}{2}L)\le t\le\tfrac{1}{2}DL\\ 0, & t \text{ otherwise.}\end{cases} \tag{5}$$

X is the short term spectrum, computed at the slow sampling rate T for frequency mF and time kT. $\mathfrak{F}$ and $\mathfrak{F}^{-1}$ are the discrete Fourier transform (DFT) and its inverse defined by the relations

$$\mathfrak{F}\{w(kT-nD)x(nD)\}=\sum_{n=kT/D-(1/2)N}^{kT/D+(1/2)N-1} w(kT-nD)x(nD)e^{-imnFD}\ , \tag{6}$$

$$\mathfrak{F}^{-1}\{X(mF,\ kT)\}=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{m=0}^{N-1} X(mF,\ kT)\,e^{imnFD}, & kT/D-\tfrac{1}{2}N\le n\le kT/D+\tfrac{1}{2}N-1,\\ 0; & n \text{ otherwise.}\end{cases} \tag{7}$$

F is the frequency sample period and is directly related to the transform length N by the relation

$$F = 2\pi/ND \ . \tag{8}$$

$w(nD)$ is the low-pass Hamming window defined by (5), and $x(nD)$ is the input signal being analyzed.

The block diagram of Fig. 2 shows two filter banks. Thus we must define a short-term spectrum for each input signal. The $y(nD)$ short-term spectrum is then

$$Y(mF, kT) = \mathfrak{F}[w(nD - kT)y(nD)]. \tag{9}$$

The operations of cophasing, adding, and scaling by G defines an output short-term spectrum

$$\hat{S} = (Y + AX)G, \tag{10}$$

where $\hat{S}, G, Y, A,$ and X are all short-term spectra defined over frequency and time in the same way Y and X are defined. G is a short-term correlation function and A is a short-term all-pass correction function. G and A will be defined in the next section.

Once $\hat{S}$ has been found, the synthesis rule[8,9] is used to find $\hat{s}(nD)$

$$\hat{s}(nD) = \sum_{k=-\infty}^{\infty} \mathfrak{F}^{-1}[\hat{S}(mF, kT)]. \tag{11}$$

Because modifications have been made to the short-term spectrum of x and y, the length of the DFT, N, must be greater than the window length and must be care fully chosen prior to the analysis step Eq. (4) in order to avoid time aliasing during synthesis. For a more detailed description of this problem, see Refs. 8 and 9.

III. DEFINITION OF G AND A

In order to define G it is necessary to define several time averages

$$\Phi_{xx}(mF, kT) = \overline{|X(mF, kT)|^2}, \tag{12}$$

$$\Phi_{yy}(mF, kT) = \overline{|Y(mF, kT)|^2}, \tag{13}$$

$$\Phi_{xy}(mF, kT) = \overline{Y(mF, kT)X^*(mF, kT)}, \tag{14}$$

where the bar means that a moving average has been performed with respect to time. X^* is the complex conjugate of X.

In our simulations the above averages were performed with an exponential weighting because then the formulas reduce to a simple recursive calculation

$$\Phi_Q(mF, kT) = \alpha\Phi(mF, (k-1)T) + Q(mF, kT), \tag{15}$$

where Q represents the term under the bar in (12), (13), and (14). α is a number close to one and defines a $1/e$ average time given by

$$\tau = -T/\ln(\alpha). \tag{16}$$

In practice τ is chosen to be frequency dependent. From (15) we see that it is a trivial extension to make α dependent on m, the frequency.

Φ_{xx} is, of course, the energy in frequency bands as a function of time for signal x. Φ_{yy} is the same for y. Φ_{xy} is the short-term cross correlation in frequency bands. In a digital system when computing second-order functions the original samples should be oversampled by a factor of two in order to avoid frequency aliasing. For the energy functions this is probably not important. For the cross-correlation function it might be. Usually this increase in the sample rate is ignored because of the nature of the signals and because of the averaging. In our simulations we did not decrease T the sample period.

We may now directly define A and G:

$$A = YX^*/|X|\ |Y| \tag{17}$$

$$G = |\Phi_{yx}|/(\Phi_{xx} + \Phi_{yy}). \tag{18}$$

Because of the definition of (17) AX will be a complex number with magnitude of $|X|$ and phase of Y. In this way $AX + Y$ averages the magnitude of complex vectors X and Y. The gain function G has the following properties. If y is only a delayed version of x with delay less than the window length L, then X^*Y will be a complex number with linear phase and G will be $\frac{1}{2}$. When the intersignal delay is greater than the window length the phase becomes random and Φ_{yx} goes to zero due to the time averaging. Thus G will be zero.

Since the operations of A and G on X and Y are in the frequency domain they act like a filter in the time domain. In order to reduce time aliasing during synthesis, a sufficiently large number of filters needs to be chosen at the outset of the analysis procedure as discussed in Ref. 8. In practice the following parameter values were used.

$$\begin{aligned} D &= 10^{-4}\ \text{sec}, \\ L &= 64, \\ T/D &= 16, \\ N &= 128. \end{aligned} \tag{19}$$

With these values the frequency spacing $F/2\pi$ is 78.13 Hz. The filter bandwidth is 625 Hz indicating a considerable filter overlap.

Other definitions for A and G are possible since they have not been derived from any rigorous theory. Of several tried by the authors, these definitions seemed to perform best. However, the question of optimal definitions of A and G is not closed at this point (e.g., an alternative definition of G is the coherence function $|\Phi_{xy}|/(\Phi_{xx}\Phi_{yy})^{1/2}$).

IV. RESULTS

The method discussed here was simulated on a digital computer. First, binaural recordings were made in reverberant rooms which had reverberation times ranging from 0.1 to 2 sec. Each recording was made by placing small electret microphones in a subject's ears and making a stereo recording on a high quality stereo tape recorder. The analogue recording was then digitized with a 12-bit analog to digital converter and stored digitally. The signals were processed on a Data General S/200 Eclipse digital computer. Processing time for 2.5 sec of speech was about 1.5 min. After processing the microphone signals, the processed speech could be direct-

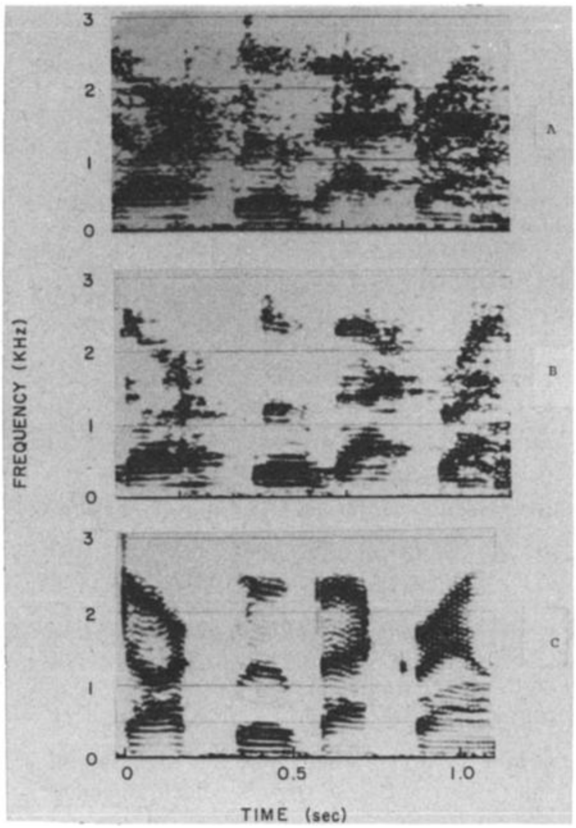

FIG. 3. Speech spectrograms: (a) reverberant, (b) processed, and (c) anechoic (separate recording).

ly compared to the original reverberant signals. For the cases of serious reverberation, (reverberation times greater than 0.5 sec) the reverberation reduction due to the processing was dramatic. In Fig. 3 we show narrow-band spectrograms of the original reverberant speech (from one channel) and the processed speech. One can see the reduction of the tails of reverberant decay which are caused by the room. The regions of suppression occur where the two microphone signals are uncorrelated. The third spectrogram was a different utterance of the same sentence by the same speaker under anechoic conditions.

V. SUMMARY

We have discussed a processing method which has been found to be very effective in removing reverberation. This method consists of a cophase and add process in bands and gain switching in bands which depends on the degree of correlation between the microphone signals in each band. The operations are equivalent to a linear time-varying filter having two input ports and one output port with the filter properties depending on the short-term spectra of the two input signals.[9]

The system was tested using real reverberation and was found to be quite effective in removing the long-term echo from the speech. Casual listening tests indicated that the processed speech was preferred over dichotically presented samples. Thus it appeared that the processing described was superior to that of the binaural hearing system.

ACKNOWLEDGMENTS

We would like to thank J. E. West and D. J. MacLean for making the recordings and M. M. Sondhi, M. R. Schroeder, and J. L. Flanagan for comments and useful criticism.

[a] Present address: Ruhr-University, 4630 Bochum, Germany F.R.

[1] J. P. A. Lochner and J. Burger, "The Intelligibility of Speech under Reverberant Conditions," Acustica 7, 195–200 (1961).

[2] Anna K. Nabelek and J. M. Pickett, "Monaural and Binaural Speech Perception Through Hearing Aids under Noise and Reverberation with Normal and Hearing-Impaired Listeners," Speech Hear. Res. 17, 724–739 (1974).

[3] Anna K. Nabelek and J. M. Pickett, "Reception of Consonants in a Classroom as affected by Monaural and Binaural Listening, Noise, Reverberation and Hearing Aids," J. Acoust. Soc. Am. 56, 628–639 (1974).

[4] J. L. Flanagan and R. C. Lummis, "Signal Processing to Reduce Multipath Distortion in Small Rooms," J. Acoust. Soc. Am. 47, 1475–1481 (1970).

[5] D. A. Berkley and O. M. M. Mitchell, "Seeking the Ideal in 'Hands-Free' Telephony," Bell Lab. Rec. 52, 318–325 (1974).

[6] A. H. Koenig, J. B. Allen, D. A. Berkley, and T. H. Curtis, "Determination of Masking Level Differences in an Reverberant Environment," J. Acoust. Soc. Am. 61, 1374–1376(L) (1977).

[7] L. Danilenko, "Binaurales Hören im nichstationaren diffusen Schallfeld," Doctoral dissertation (University of Aachen, FDR, 1968).

[8] Jont B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. Acoust. Speech Signal Process. ASSP-25, 235–238 (1977).

[9] Jont B. Allen and L. R. Rabiner, "A Unified Approach to short-time Fourier Analysis and Synthesis," Proc. IEEE 65, 11 (1977).