SPEECH DEREVERBERATION: PERFORMANCE OF SIGNAL PROCESSING ALGORITHMS AND THEIR EFFECTS ON INTELLIGIBILITY

P. JEFFREY BLOOM AND G. D. CAIN

DIVISION OF ENGINEERING, POLYTECHNIC OF CENTRAL LONDON

ABSTRACT: Results of a previous evaluation of a two-input signal processing dereverberation technique [1] indicated that, on average, the intelligibility of isolated reverberated words was not significantly altered by processing, even though a decrease in measured and perceived reverberation time was observed [3]. In this paper, we offer explanations for these earlier findings based on a detailed analysis of the following critical factors: 1) The effect of the room impulse response on the short term speech spectrum; 2) The problems associated with using the frequency-dependent interaural coherence estimate directly as a speech-filter gain modulator; 3) The influence of the expected direct-to-reflected energy ratio on the maximum expected interaural coherence; and 4) The effect that the processors' "noise" suppression has on the perceptually important acoustic features of the speech waveform. Accordingly we have made experimental modifications to the original processing techniques which are expected to improve its performance. Demonstrations of the modified processes' effects on relevant test signals and on word identification are presented and compared to those of the original process. Lastly, it is shown how many of the observations and operations associated with dereverberation processing are directly applicable to the more general problem of intelligibility enhancement in noisy environments.

I. A Model of One Source and Two Receivers in Small Reverberant Rooms

In this paper, we are considering the following physical model. A single point source emits a signal s(t) into a room in which two spatially separated microphones -- both located a distance d from the source -- pick up signals x(t) and y(t) respectively. The transmisson path between the source and, say, the receiver detecting x(t) may be described by a room impulse response $h_1(t)$. One such typical response is shown schematically in Fig.1. The signal x(t) may, therefore, be found mathematically by forming the convolution $x(t)=s(t)*h_1(t)$. Alternatively, we may note from experimental evidence, that $h_1(t)$ is composed of a directly-received impulse, which we will call $p_1(t)$ and a reverberant "tail" which will be called $g_1(t)$, made up entirely of reflections. Therefore,

$$x(t) = s(t)*p_1(t) + s(t)*g_1(t) . \qquad (1)$$

Similarly, $y(t) = s(t)*p_2(t) + s(t)*g_2(t)$ so that the complete "system" can be modeled as in Fig.2. x and y can represent the signals at a listener's ears.

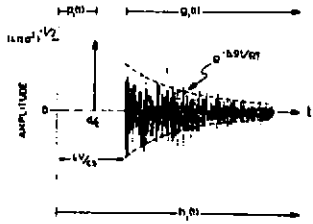

Fig.1: Room Impulse Model

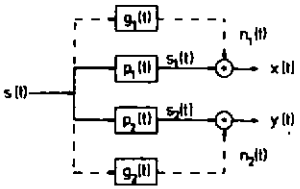

Fig.2: Room System Model

For a given room and fixed d, $g_i(t)$ will be vastly different for every change in source or receiver positions [2]; this leads us to model g(t) as a non-stationary random process. Specifically, let $g_i(t)$ be a sample member of the random process

$$\underline{g}(t)=I^{\frac{1}{2}}\exp(-mt)\ \underline{w}(t)\ u(t-\Delta t) \qquad (2)$$

where I is a constant to be determined; $\underline{w}(t)$ is stationary, bandlimited, white noise with zero mean and unit variance; and $u(t-\Delta t)$ is a unit step function which "turns on" when the first reflection is received at Δt. Although our model for $\underline{g}(t)$ does not properly describe the first few discrete reflections, normalized autocorrelation functions (NACF) of real and simulated room impulses indicate that NACF ≤ 0.1 for lags slightly greater than 0, which is characteristic of "white" processes. It is also well known [2] that $\langle g^2(t)\rangle \propto \exp(-13.8|t|/RT)$ where RT is the usual (60 dB) definition of reverberation time and $\langle\ \rangle$ indicates time averaging; therefore the value of m in (2) is easily identified as $m = 6.9/RT$.

Assuming a uniform directivity pattern for both the source and receivers, with c being the speed of sound in air, we can define the direct path "filters" as

$$p_1(t) = p_2(t) = p(t) = \delta(t-d/c)/(4\pi d^2)^{\frac{1}{2}}. \qquad (3)$$

II. A Link Between the Direct-to-Reflected Energy Ratio and Coherence

In a limiting case of a source with continuous power generating a diffuse sound field, the average direct-to-reflected energy ratio at a fixed distance d is [2]

$$k = -S\ln(1-\alpha)/16\pi d^2(1-\alpha) \ . \qquad (4)$$

Here, S is the total surface area of the room, and α is the average absorption coefficient. It will emerge that this ratio k (or a time-varying version) is critical to the operation of any two-channel speech dereverberation processing.

We now introduce the usual definition [4] for magnitude coherence $C_{xy}(f)$ between two signals x(t) and y(t):

$$C_{xy}(f) = |\Phi_{xy}(f)|/[\Phi_x(f)\ \Phi_y(f)]^{\frac{1}{2}} \qquad (5)$$

where the power spectral density function $\Phi_x(f) \triangleq \int_{-\infty}^{\infty} R_x(t')\exp(-j2\pi ft')\ dt'$ and the autocorrelation function $R_x(t') \triangleq E\,[x(t)\ x(t+t')]$ (with E being the expectation operator); $\Phi_y(f)$ and $\Phi_{xy}(f)$ are defined similarly. From here on, we will suppress any explicit dependence on f; hence $C_{xy} = C_{xy}(f)$. As discussed in [1] and [4], this function has extremely useful properties, which in our application will shortly be evident.

If s(t) in Fig.2 is a zero-mean, stationary white noise, with average intensity given by $R_s(t') = I_o\delta(t')$, we can substitute the expressions for $g_i(t)$ and p(t) (Eqs.2,3) into this room model (Eq.1) and derive the following relations:

$$R_x(t') = R_y(t') = [I_o\,\delta(t')/(4\pi d^2)] + [RT\ I_o I\delta(t')/13.8] \qquad (6)$$

$$R_{xy}(t') = R_{s_1s_2}(t') = I_o\delta(t')/(4\pi d^2) \qquad (7)$$

Substituting the Fourier transforms of (6) and (7) into (5), we find

$$C_{xy}(f) = [1 + (RT/13.8)\ I\ \exp(-13.8\Delta t/RT)\ 4\pi d^2] = 1/(1+k^{-1}) \qquad (8)$$

where the final form is obtained by substituting the mean free path length, $4V/S$ for Δt_c, and $-0.163V/S\ln(1-\alpha)$ for RT [2]. The relationship, $I = 125/V$ is derived by arguing that the average ratio of the direct-to-reverberant power in narrow bands of frequencies (found approximately from $\Phi_{s_i}(f)/\Phi_{n_i}(f)$) must be the same for all f and thus equal to k in (4). The important relationship in

(8) clearly shows the effect that k has on the magnitude coherence. If there is no direct signal k=0, and $C_{xy}=0$; if k is very large (no reverberation) $C_{xy}=1$; for $0<k<\infty$, $0<C_{xy}<1$. Similar effects to these were informally suggested in [1]. Finally, we can rearrange (8) to find [4]

$$\Phi_{s_i}/\Phi_{n_i} = k = C_{xy}/(1-C_{xy}). \qquad (9)$$

We now see that under the conditions imposed by our model, the direct-to-reverberant power ratio vs. frequency may be found by measuring $C_{xy}(f)$. (Note however that when the wavelength of the signal is greater than roughly twice the receiver separation, C_{xy} will always tend towards 1.0.)

The next question, of course, is how (9) can be used. One interesting approach with a familiar result is to note that a minimum mean-squared estimate of a signal s(t) corrupted by noise n(t) when x(t) = s(t)+n(t) can be found by forming the signal estimate $\hat{S}(f) = [\Phi_s' /(\Phi_s' + \Phi_n')] X(f)$. Here the ' means a smoothed spectral estimate and X(f) is a short time Fourier transform of the (windowed) input. Simple rearrangement of this equation and substitution of (9) as an estimate of Φ_s' /Φ_n' yields:

$$\hat{S}(f) = C_{xy}(f)\, X(f) \qquad (10)$$

Thus by taking advantage of having <u>two</u> inputs, each containing the signal and some uncorrelated noise, we can use the magnitude coherence function <u>directly</u> to modify optimally, in the mean square error sense, one (or possibly both) input(s). This is exactly the approach suggested in [1] and used in [3] for dereverberation (with the minor exception that geometric average of power in the denominator of (5) was replaced by an arithmetic averaging -- making only a constant factor of 2 difference when input channel powers are equal). We note here, however, that <u>any</u> additive noise that is uncorrelated between the receivers and contributes to both n_1 and n_2 will be suppressed in the same manner as the reverberant signal. Thus, such a system may be useful in noisy environments where no <u>direct</u> noise component is detected in both channels.

III. <u>Consideration of the Time-varying Nature of Speech, k, and $C_{xy}(f)$</u>

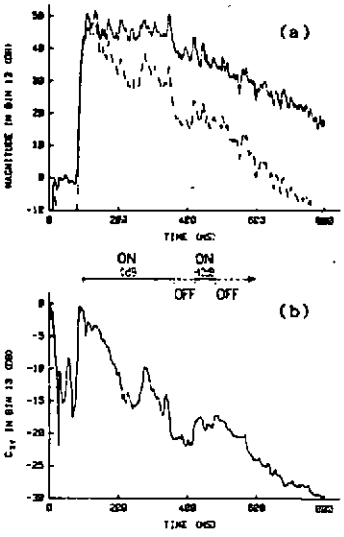

Fig.3: Processing Experiment

Speech recognition depends critically on detecting certain time-varying acoustic features which provide cues to phoneme identity [5]. Reverberation, because of its nature of producing "noise" whose power and spectra are related closely to those of a signal, tends to often degrade the recognition of stop consonants which follow higher intensity sounds such as vowels [3]. We can illustrate the underlying reason for this with an experiment in which a test signal, with speech-like temporal properties, was played into a room (with RT = 1.0 s). The resulting reverberant signal was recorded binaurally (d=4m) and subjected to the dereverberation processing [3]. The test signal comprised lowpass filtered (f ≤ 2.5 kHZ), gated random noise which was "on" for 275 ms with relative amplitude 0dB, off for 50ms, and "on" again for 62ms with relative amplitude -12dB. Such temporal/amplitude relationships are roughly typical of those found in mid- to high-frequency bands during vowel-stop C utterances. This on/off (at -6dB) pattern is indicated between Fig.3 a and b. In Fig.3a the output of one DFT bin during short time analysis is plotted vs. time, and is "smoothed"

slightly (to remove distracting fine structure). The burst of the test pattern has a well-preserved leading edge and maintains constant amplitude during the burst; but during the off period, the room retains signal energy, which falls off as exp(-13.8t/RT). By the start of the second, lower-amplitude burst, reverberant energy masks both the leading edge and the burst itself. After the second burst has stopped the room decay is clearly seen. In Fig.3b we show the measured value of C_{xy} in the same frequency bin. C_{xy} immediately rises to 1.0 when the initial transient is detected, but since the direct signal is constant, the reverberant energy increases expotentially; thus, k, as given in (9) decreases expotentially and C_{xy} must fall according to (8). Note that the value of k calculated for stationary conditions will provide a "target" or limiting value when the source produces (as in this example) constant power. When the first burst stops, only reverbrant energy remains and C_{xy} drops suddenly (theoretically to $-\infty$, but limited here by experimental noise).

At the onset of the second burst, the leading edge is clearly detected in C_{xy}, but in the presence of the reverberation from the first burst, so the resulting values of C_{xy} remain relatively low. At the end of the second burst, C_{xy} drops below the theoretical limit for stationary C_{xy}, again decaying slowly due to experimental noise. In summary, we see that C_{xy} will tend to fall towards a limiting value (as set out in (8)) when the signal output is constant; however, C_{xy} will also fall (but at a higher rate) when it is above this limit and the direct signal decreases rapidly in amplitude. The first effect can cause undesirable signal modulation, whereas the second effect will tend to remove the reverberant portion of a signal. The result of modulating the input signal in Fig.3a (solid line) by C_{xy} in Fig.3b is shown in Fig.3a (broken line), and illustrates these effects.

IV. Suggestions for Processing Improvement

It is our contention that the processing scheme suggested in [1] is unlikely to enhance intelligibility, partly for the reasons explained above. It would appear, however, that the time-varying coherence function may be of more use as an input to a second stage of signal detection and processing. In such a second stage, gain functions can be devised which attempt to "boost" or enhance leading "edges" of the signal (in a particular frequency band) when C_{xy}, measured in that band, has either exceeded a selectable threshold, or increased at a rate exceeding a threshold rate. The thresholds can be parameters adjusted to suit the particular room and distance conditions in which the processor was operating. By emphasizing those aspects of the signal which seem to carry perceptually significant acoustical features, it may be possible to improve the recognition of certain otherwise unintelligible phonemes.

Acknowledgement This work was supported by the Medical Research Council (Grant no. G978/607/N), which the authors gratefully acknowledge.

References

1. J.B.ALLEN, D.A.BERKLEY and J.BLAUERT 1977 J.Acoust.Soc.Am. 62 912-915. Multimicrophone signal-processing technique to remove reverberation from speech signals.
2. H.KUTRUFF 1979 Applied Science Publishers London Room Acoustics, 2nd ed.
3. P.J. BLOOM 1980 IEEE ICASSP-80, Denver, Colorado, April 9-11, Evaluation of a Dereverberation Process by Normal and Impaired Listeners.
4. G.C.CARTER, C.H.KNAPP and A.H.NUTTALL 1973 IEEE Trans.Aud.Electroacoustics AU-21, 337-344. Estimation of the Magnitude-Squared Coherence Function via Overlapped Fast Fourier Transform Processing.
5. K.N.STEVENS 1980 J.Acoust.Soc.Am. 68 836-842. Acoustic correlates of some phonetic categories.