THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 8, October 1981
Printed in U.S.A.

Improving the Quality of a Noisy Speech Signal

By M. M. SONDHI, C. E. SCHMIDT, and L. R. RABINER

(Manuscript received December 18, 1980)

In this paper we discuss the problem of reducing the noise level of a noisy speech signal. Several variants of the well-known class of "spectral subtraction" techniques are described. The basic implementation consists of a channel vocoder in which both the noise spectral level and the overall (signal + noise) spectral level are estimated in each channel, and the gain of each channel is adjusted on the basis of the relative noise level in that channel. Two improvements over previously known techniques have been studied. One is a noise level estimator based on a slowly varying, adaptive noise-level histogram. The other is a nonlinear smoother based on inter-channel continuity constraints for eliminating the so-called "musical tones" (i.e., narrowband noise bursts of varying pitch). Informal listening indicates that for modest signal-to-noise ratios (greater than about 8 dB) substantial noise reduction is achieved with little degradation of the speech quality.

I. INTRODUCTION

The idea that a vocoder may be used to improve the quality of a noisy speech signal, has been around for about twenty years. To the best of our knowledge the first such proposal was made in 1960 by M. R. Schroeder.[1] The basic idea of this proposal can be explained with the help of Fig. 1, as follows:

Figure 1a shows a typical short-term magnitude spectrum of a voiced portion of a noisy speech signal. Let $S(\omega)$ denote the envelope of this spectrum. (Recall that the "channel gains" of a vocoder are estimates of this envelope at the center frequencies of the channels. The fine structure of the spectrum is attributed to the harmonics of the fundamental voice frequency.)

Figure 1b shows a "formant equalized" version, $\bar{S}(\omega)$, of the envelope. The peaks in S and $\bar{S}$ occur at the same frequencies but the peaks of $\bar{S}$ (unlike those of S) are all of the same amplitude.

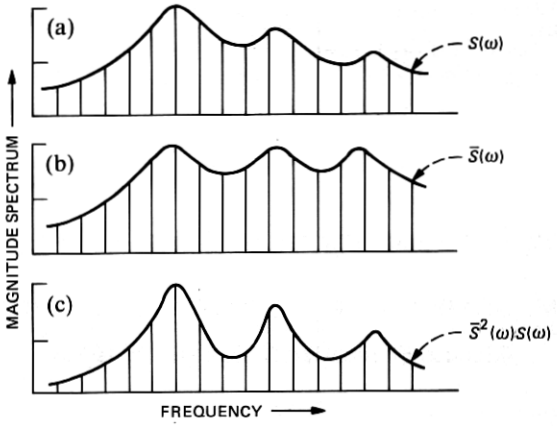

Fig. 1—Illustration of noise stripping by increasing the dynamic range between formant peaks and noise valleys. (a) Original spectral envelope and fine structure. (b) Formant-level equalized spectral envelope. (c) The product spectrum $\bar{S}^2(\omega)S(\omega)$ in which the ratios between formant peaks and valleys is larger than in the original spectrum.

The proposal is, essentially, to generate a signal with a fine structure as close as possible to that of the original speech signal, but with an envelope given by $\bar{S}^n S$, where n is some intetger, say, 1 or 2. Except for a scale factor, the spectral envelope of the resulting signal is the same as that of the original signal at the formant peaks, but is considerably reduced in the valleys. As shown in Fig. 1c this processing effectively reduces the overall noise level. Of course, the formant peaks also become sharper, i.e., the formant bandwidths get reduced.

Reference 1 describes two implementations of this idea: a frequency domain method in which the envelope is modified by modifying the channel gains of a self-excited channel vocoder, and a time domain method in which the same effect is achieved by repeated convolution.

In many practical cases of interest, the noise is additive and uncorrelated with the speech signal. In such a situation, if it were possible to estimate the spectral level of the noise as a function of frequency, then the noise reduction could be achieved in a somewhat different manner. Suppose the noisy speech is applied to the input of a channel vocoder (see Section II for a detailed description). Let the output of the kth channel be $y_k = s_k + n_k$, where s_k is in the speech signal and n_k the noise signal in that channel. Let N_k^2 be the average power of the noise and S_k^2 that of the speech signal. Then, assuming that the noise and speech are uncorrelated, the average power of the noisy speech is given by

$$Y_k^2 = S_k^2 + N_k^2 \tag{1}$$

Now Y_k^2 can be estimated directly from the output signal y_k. If an

estimate of N_k^2 is available, as postulated, then $(Y_k^2 - N_k^2)^{1/2}$ provides an estimate of the magnitude of the signal alone in the kth channel. Thus, if the level of the channel signal is multiplied by the ratio of this estimated signal power to overall power, then a noise reduction is achieved.

In 1964, at the suggestion of M. R. Schroeder, this "spectral subtraction" idea was implemented as a BLODI language computer program by one of us (MMS) in collaboration with Sally Sievers.[2] Besides spectral subtraction, one other feature was incorporated into this implementation. It had been recently demonstrated that autocorrelation and cepstrum pitch extraction are quite accurate and reliable for noisy speech signals with signal-to-noise ratio (s/n) as low as 6 dB.[3,4] Such extractors provide a clean excitation signal even from a highly noisy speech signal. Therefore, the self-excitation described in Ref. 1 was replaced by a voiced-unvoiced (buzz-hiss) signal derived from an autocorrelation pitch extractor.

Although this implementation demonstrated the feasibility of the basic idea, the computer facilities available at that time did not allow a thorough investigation of the effects of changing various parameters and configurations. Also, since digital hardware was not yet readily available, it did not appear likely that such noise-stripping techniques would find application in the immediate future. For these reasons these techniques were not actively pursued at that time.

Since the mid-seventies, presumably due to the vastly improved digital technology and renewed military interest, noise-stripping has again attracted considerable attention. The renewed interest in this problem appears to have started in 1974, when Weiss et al. independently discovered the spectral subtraction method.[5] Except for the fact that the filter bank of the channel vocoder was replaced by short-term Fourier analysis, the implementation of Weiss et al. was quite similar to the one described above. During the past five or six years several studies have explored this and other methods for noise removal. Notable among these is the work of Boll, Berouti et al., and McAulay and Malpass.[6,7,8] A review of these and other studies is given in a recent paper by Lim and Oppenheim.[9]

In view of the current interest in noise removal, we have recently been experimenting with the spectral subtraction method by computer simulation. Subsequent sections of this paper describe the results of our experiments.

From the brief description given above, it is clear that spectral subtraction is expected to be useful only in cases when the noise is additive. With this constraint, there are basically two types of situations in which this method might find application:

(*i*) The speech may be produced in a noisy environment, e.g., in

the cockpit of an airplane. In such a situation the spectrum of the noise is unknown a priori. This information must be estimated from the noisy speech signal itself, e.g., during intervals of silence between speech bursts. The algorithm for estimating the noise spectrum is, therefore, one of the most important parts of the simulations described later.

(*ii*) The speech itself may be generated in a quiet environment but might be transformed to a noisy signal because of the action of a coding device. Examples where such noise may be modelled as additive are pulse-code modulation (PCM) coders, and delta modulators whose step size is chosen such that granular noise predominates over the slope-overload noise. In such cases, both the level of the noise and its spectral composition might be known a priori. Use of this a priori information simplifies the system and improves its performance.

There is a third way in which noise may enter the communication channel additively. The speech signal may be generated in a quiet environment but the listener may be in a noisy environment. A message sent over the public address system at a busy railway station is such an example. In this case, the problem is to preprocess the speech signal in such a way that its intelligibility is least impaired by the noise. Some work on this problem has been reported in the literature;[10] however, we will not deal with this problem.

Before turning to a description of our simulations, it is worth emphasizing that we deliberately used the word "quality" rather than "intelligibility" in the title of this paper. Ideally, of course, one would like the intelligibility also to be increased. However, this is not absolutely essential. It is quite annoying and fatiguing to have to listen to a noisy speech signal for any length of time. Therefore, a device that reduces or eliminates the noise can be quite useful even if the cleaner signal is no more intelligible than the noisy one.

II. THE BASIC STRUCTURES

Two basic channel vocoder configurations for implementing spectral subtraction were simulated. For reasons that will become apparent from the following descriptions, we call these configurations self-excited and pitch-excited, respectively.

2.1 The self-excited configuration

A block diagram of the self-excited method of noise removal is shown in Fig. 2. The noisy speech, sampled 10,000 times per second is first passed through a bank of N equispaced bandpass filters that span the telephone channel bandwidth (approximately 200 to 3200 Hz). The processing of the output of the bandpass filter is identical for each

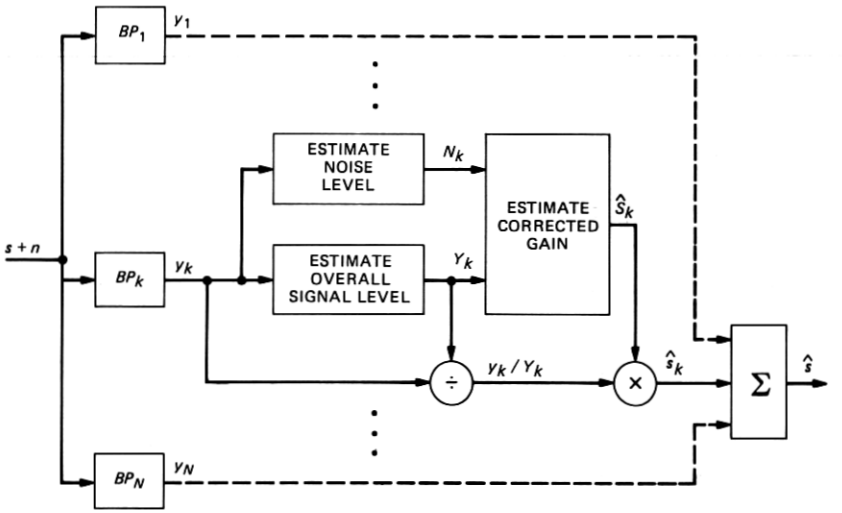

Fig. 2—Block diagram of the self-excited channel bank noise stripper consisting of a bank of N FIR bandpass filters with gain estimation and correction within each channel.

channel. In the kth channel, the following operations are performed on the output y_k:

(*i*) The level (magnitude) of the noisy speech signal, Y_k, is estimated.

(*ii*) In a parallel path the level of the noise, N_k, is estimated.

(*iii*) The estimates N_k and Y_k are used to derive an estimate $\hat{S}_k$ of the level of the uncorrupted speech signal in the kth channel.

(*iv*) The adjusted channel signal is computed by the relation

$$\hat{s}_k = y_k \frac{\hat{S}_k}{Y_k}. \tag{2}$$

Clearly $\hat{s}_k$ has the desired estimated magnitude $\hat{S}_k$. The sum $\hat{s} = \sum_{k=1}^{N} \hat{s}_k$ then provides the final processed output.

2.2 The pitch-excited configuration

A block diagram of the pitch-excited method is shown in Fig. 3. The estimates $\hat{S}_k$, $k = 1, 2, \cdots N$, are obtained exactly as in the case of the self-excited configuration. However, the adjusted channel signals are obtained differently.

(*i*) The noisy speech signal is first processed by a pitch extractor which also provides the voiced/unvoiced classification. The particular pitch extractor used is described in Ref. 11.

(*ii*) The output of the pitch extractor is used to provide a clean excitation signal which consists of a Gaussian noise during unvoiced

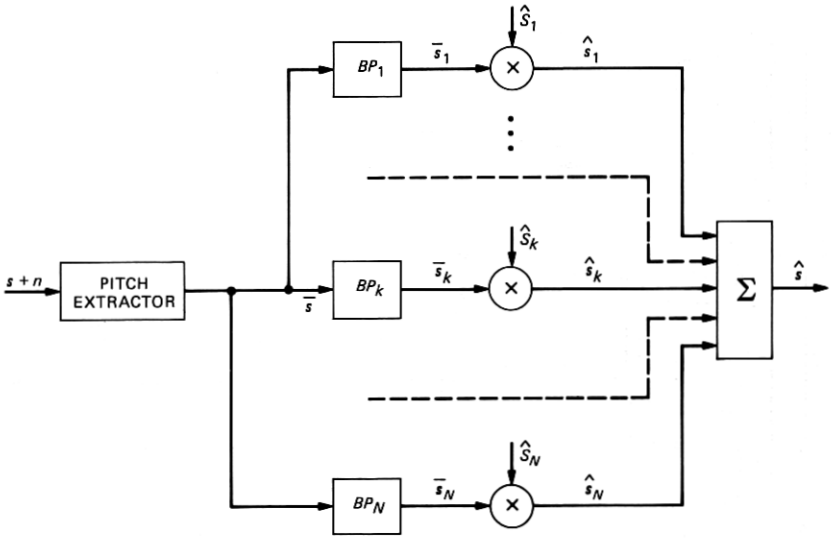

Fig. 3—Block diagram of the pitch-excited channel bank noise stripper in which a voiced/unvoiced excitation is used in place of the bandpass channel signals.

portions and a train of impulses at the pitch rate during voiced segments.

(*iii*) This clean excitation signal is passed through a bank of bandpass filters, identical to the ones shown in Fig. 2, to give channel signals, $\overline{s}_k$, which are approximately equal in magnitude.

(*iv*) The adjusted channel signal is computed as

$$\hat{s}_k = \overline{s}_k \cdot \hat{S}_k. \tag{3}$$

As before, $\hat{s}_k$ has the correct magnitude and, as before, the sum of these adjusted channel signals gives the final processed output.

As discussed in the next section, the estimates of $\hat{S}_k$ are computed every 0.01 s (i.e., 100 times a second). In our initial experiments the channel gains were held constant between estimates. In this case, the gain jumps in value every 0.01 s, producing annoying audible clicks. These clicks were eliminated by replacing each jump by a linear interpolation of the channel gains over 6 speech samples (i.e., over 0.6 ms).

III. ALTERNATIVE CONFIGURATIONS SIMULATED

Several modified versions of the basic configurations of Figs. 2 and 3 have been simulated, and several sentences processed with these simulations. The alternatives that we have studied in some detail are two choices for the number of channels; two methods of estimating

Y_k; two methods of estimating N_k; and two methods of estimating $\hat{S}_k$. These will now be described.

3.1 The filter bank

Two designs were simulated, each with equispaced filters. In one design 16 channels (200-Hz wide) were used, and in the other 32 channels (100-Hz wide). The filter responses and the sum of the responses for each design are shown in Fig. 4. (Each filter was a linear phase, finite impulse response (FIR) filter of duration 88 samples in the 16-channel filter bank and 176 samples in the 32-channel filter bank.)

3.2 Estimating Y_k

The two methods of estimating the magnitude, Y_k, of the noisy channel signal are shown in Fig. 5. Either $|y_k|$ or y_k^2 is low-pass-filtered to 30 Hz. In the second case, the square-root of the output of the low-pass filter is computed. The impulse and frequency responses of the low-pass filter [a 3rd order infinite impulse response (IIR) Bessel filter] are shown in Fig. 6.

The choice of bandwidth of the low-pass filter is governed by a compromise between the following two requirements: For accurate estimation of Y_k the averaging time should be as large as possible, i.e.,

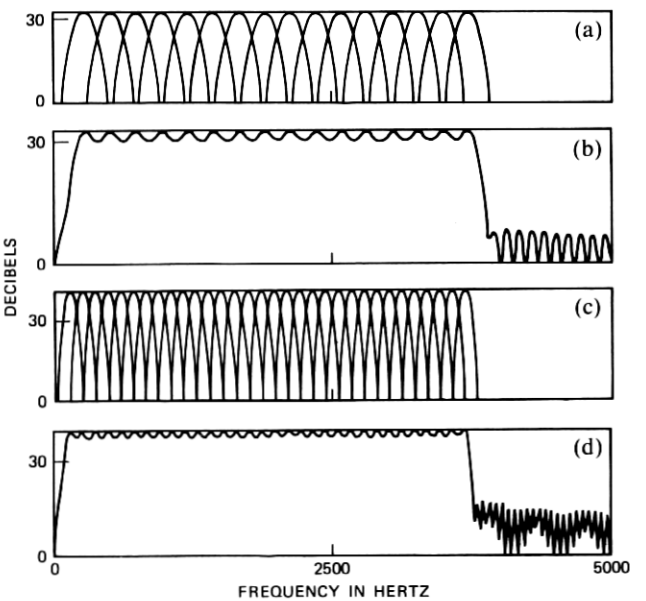

Fig. 4—(a) Frequency responses of individual filters of the 16-channel filter bank. (b) Composite responses for 16-channel filter bank. (c) Frequency responses of individual filters of the 32-channel filter bank. (d) Composite responses for 32-channel filter bank.

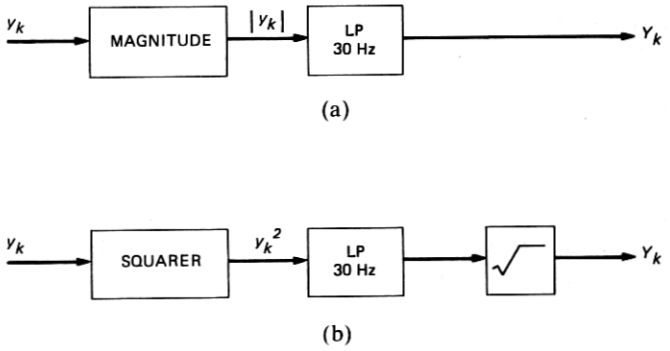

Fig. 5—Signal processing for estimating the overall signal level using either a magnitude (a) or a squaring (b) nonlinearity, followed by a low-pass filter. In the case of the squaring nonlinearity, the low-pass filter is followed by a square root box.

the filter bandwidth should be as small as possible. On the other hand, the spectrum of speech varies with time so the bandwidth should be as large as possible to track these variations. The usual compromise cut-off frequency in channel vocoders is about 30 Hz.

Note that the outputs of the low-pass filters need be sampled only 60 times/s. To allow for the roll-off of the filters, the sampling rate was chosen as 100/s. Somewhat surprisingly, a much higher sampling rate was found to degrade performance. We will explain this paradox in Section IV (The Musical Tones).

3.3 Estimating N_k

During intervals of silence in the speech, the input signal consists of noise alone. Therefore, one possible estimate for N_k is the smallest value attained by Y_k. However, because of statistical fluctuations, Y_k quite rapidly takes on an unrealistically low value. Therefore, this estimate is quite unsatisfactory. In order to avoid such problems with outliers, the method schematized in Figure 7 has been simulated.

As a first step, the magnitude of y_k is estimated by a procedure identical to that of Fig. 5, except that the low-pass filter has a cut-off frequency of 10 Hz instead of 30 Hz. (The impulse response of the 10-Hz filter is quite similar to that of the 30-Hz filter with the time axis scaled by a factor of 3.)

As before, the cut-off frequency of the low-pass filter should be chosen no larger than that necessary to follow the time-variations of the noise spectrum. Our choice of 10 Hz is an extremely conservative value. For most applications a cut-off frequency of 1 Hz or less should suffice.

Analogously to the estimation of Y_k, we have two ways of estimating

N_k, which differ only in the type of nonlinearity used. Figure 7 shows the front end of the alternate noise estimator that we have simulated.

Let $Z_k(n)$ be the estimates of the magnitude of y_k obtained by one of these methods, sampled every 0.01 s. Then the algorithm for finding the noise level is as follows:

(*i*) Store $Z_k(n)$, $n = 1, \cdots, Q$ in a buffer of size Q.

(*ii*) Find the smallest value such that the next higher value is within 6 dB of it. Call this smallest value MIN.

(*iii*) Make a histogram with 1-dB bins of all the values that lie in the range MIN to MAX = MIN + 15 dB.

(*vi*) Declare K times the magnitude corresponding to the peak of the histogram, as the noise level.

(*v*) Get next sample.

(*vi*) If this sample is greater than MAX, discard it and go to step (*v*).

(*vii*) If the sample is less than MAX replace the oldest sample in the buffer by the new sample and go to step (*ii*).

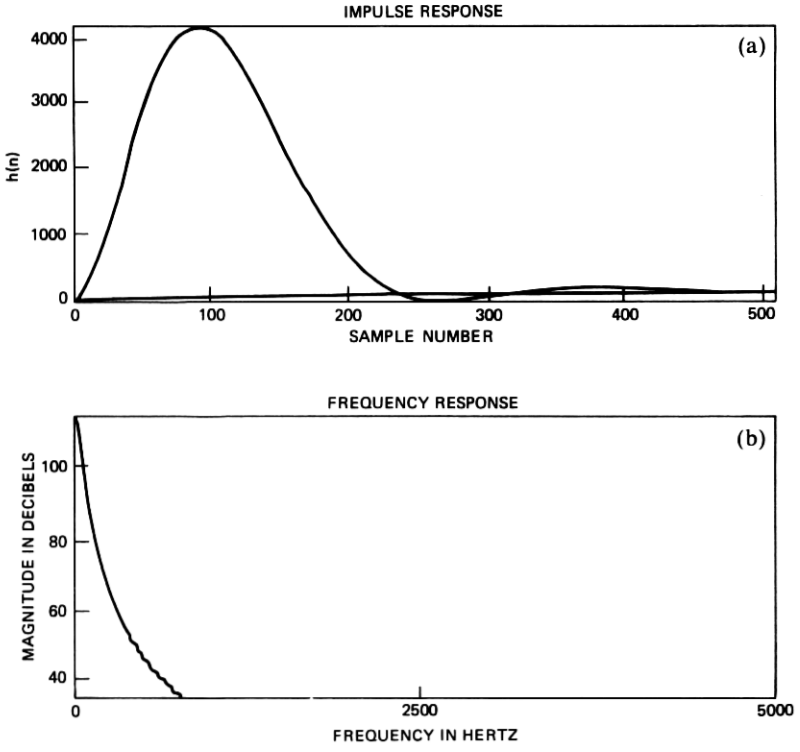

Fig. 6—Impulse response (a) and frequency response (b) of the 30-Hz, 3rd order, Bessel IIR filter used in estimating overall signal level.

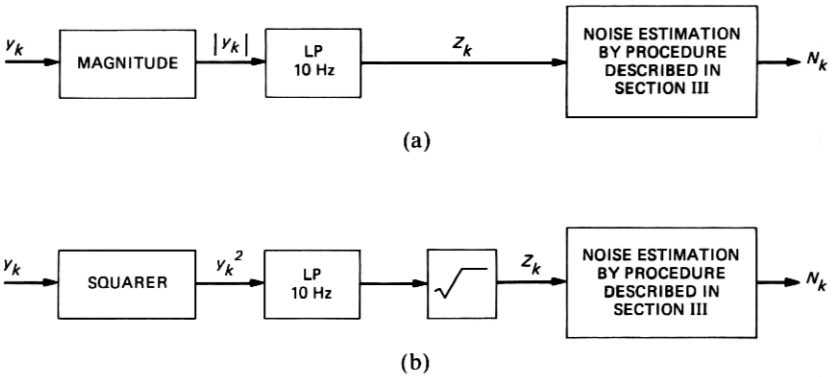

Fig. 7—Signal processing for estimating noise level. In (a) and (b) the estimates of channel levels are obtained exactly as in (a) and (b) of Fig. 6, except that the low-pass filters have a bandwidth of 10 Hz instead of 30 Hz. The final step in both (a) and (b) is an adaptive noise estimation procedure based on a time-varying noise histogram.

After some experimentation, $Q = 100$ and $K = 3$ or 3.5 were found to be most satisfactory for the range of s/n's considered. All experiments to be described later were performed with these values of Q and K.

Careful considerations of the above algorithm should convince the reader that this procedure ignores occasional low values of Z_k; it guards against sudden increased in Z_k because of the onset of speech; and finally, it allows adaptation to a slowly varying noise level.

3.4 Estimating $\hat{S}_k$

As mentioned in the introduction, under the assumption that s_k and n_k are uncorrelated, $\hat{S}_k$ should be estimated as $\hat{S}_k = (Y_k^2 - N_k^2)^{1/2}$. However, there is statistical fluctuation because of the finite averaging time even if the assumption is strictly valid. Therefore, sometimes the estimated value of Y_k is less than that of N_k. In such cases, $\hat{S}_k$ is set to zero. Thus, our first procedure for estimating $\hat{S}_k$ is

$$\hat{S}_k = \sqrt{Y_k^2 - N_k^2}, \qquad Y_k > N_k \tag{4a}$$

$$= 0, \qquad Y_k \leq N_k. \tag{4b}$$

A second estimate that we have tried is

$$\hat{S}_k = Y_k - N_k, \qquad Y_k > N_k \tag{5a}$$

$$= 0, \qquad Y_k \leq N_k. \tag{5b}$$

IV. THE MUSICAL TONES

We have processed several speech signals through a variety of noise-stripping algorithms obtained by selecting from the alternatives listed above. The results will be discussed in detail in the next section.

However, one general observation that can be made is that although the noise can be eliminated even from severely noisy speech signals, it gets replaced by "musical tones." These are short bursts of more or less sinusoidal tones with varying pitch. The explanation of the origin of these tones is as follows: The algorithm for estimating $\hat{S}_k$ will, in general, set several consecutive channels to zero (in the valleys between formant peaks). If because of statistical fluctuation a single channel escapes elimination, (i.e., is above the noise threshold) it will appear as a narrow band signal much like a tone-burst with the center frequency of the channel. Every time such a tone-burst appears, its pitch will be determined by the particular isolated channel that gives rise to it. (At this point, the paradox mentioned in Section 3.2 can be explained. Consider a channel where the speech energy is very low, i.e., $Y_k \approx N_k$. If Y_k is oversampled, the number of times it crosses N_k increases and, therefore, the number of spurious noise bursts also increases.)

We have found one simple procedure to combat this phenomenon. Every time the channel gains $\hat{S}_k$ are updated, the new values are scanned across channels (i.e., the array $\hat{S}_k(n)$, $k = 1, \cdots N$ is examined at the time instant n). A nonzero value which is flanked by zero on both sides, is set equal to zero.

For male voices, this removal of isolated channels works extremely well. However, the method does not work well for high-pitched female voices when the noise level is high. The reason is that in the latter case there may be only one or two pitch harmonics in a formant peak. Thus, the noise stripping algorithm might create several isolated channels in formant regions as well. Therefore, removal of isolated channels removes a large part of the speech signal, along with the musical tones. We do not have a good method of dealing with this problem for high-pitched voices at high noise levels.

Suggestions for combatting these musical tones have also been made by Boll and Berouti et al.[6,7] We have compared our method to these other methods and find that except in the case of high-pitched voices at very low s/n our method performs better.

V. EXPERIMENTS

We have processed several sentences spoken by male and female speakers through noise-strippers obtained by selecting most of the possible combinations of alternatives listed in Section II. Uncorrelated Gaussian noise was added to provide the noisy test samples. The variance of the Gaussian distribution was selected so as to provide several s/n's in the range of about 4 to 16 dB. We have not conducted formal listening tests on the outputs. However, informal listening

(mostly by the three authors of this report) allows us to draw the following general conclusions:

(*i*) The algorithm is capable of following slow variations of the noise spectrum. We tested this on noise with a flat spectrum but with a sudden jump of 6 dB in its amplitude. The algorithm attained the correct estimates of channel gains within 0.5 s.

(*ii*) The implementation with the 32-channel filter bank performs better than the one with the 16-channel filter bank.

(*iii*) In Fig. 5 the second alternative performs significantly better than the first, i.e., the square root of the average power is a better statistic to use than the average of the magnitude.

(*iv*) Power subtraction [eqs. (4a, 4b)] and spectral magnitude subtraction [eqs. (5a, 5b)] appear to work about equally well even at the lowest s/n (about 5 dB) that we tried.

(*v*) The factor K in the noise estimation procedure of Section III, should be set to about 3 or 3.5 for the range of s/n's considered in this paper.

(*vi*) For male voices, if isolated channels are eliminated as discussed in Section III, then pitch excitation and self excitation both work about equally well.

(*vii*) For female voices it is not possible to remove isolated channels at high noise levels (s/n's less than say 8 dB). In these situations, pitch excitation is superior to self excitation.

VI. CONCLUSION

We have described several algorithms based on spectral subtraction for removing noise from a noisy speech signal. Two noteworthy features of our simulations are the manner in which we estimate the noise level and the manner in which we deal with the narrow-band, time-varying noise bursts that commonly arise in spectral subtraction methods.

Our simulations were arranged to provide flexibility to allow us to test various modifications. However, it should be possible to realize the final preferred version of our algorithm in digital hardware that runs in real time.

The ultimate test of such a system is a large-scale statistical study of listeners' preference. We have not attempted such a study. However, on the basis of informal listening we can say that our method is quite successful in removing noise, and in most instances is superior to the other methods known to us.

REFERENCES

1. M. R. Schroeder, U.S. Patent No. 3,180,936 filed December 1960, issued April, 1965.

2. This simulation by M. M. Sondhi and S. Sievers is documented only in an unpublished internal report. The procedure is described briefly in a review paper by M. R. Schroeder and A. M. Noll, Paper A21, 5th Int. Congress on Acoust., Liege, Belgium, 1965. The device was also patented by M. R. Schroeder, U.S. Patent No. 3,403,224, filed May 1965, issued September, 1968.
3. M. M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. Audio, *AU-6*, No. 2 (June 1968), pp. 262–6.
4. A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am, *41*, No. 2 (February 1967), pp. 293–309.
5. M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Processing Speech Signals to Attenuate Interference," IEEE Symp. on Speech Recognition, Pittsburgh, April 1974, Contributed Papers, pp. 292–3.
6. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoust. Speech and Sig. Process., *ASSP-29*, No. 2 (April 1979), pp. 113–20.
7. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," Proc. Int. Conf. Acoust. Speech, and Sig. Process. (April 1979), pp. 208–11.
8. R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Maximum Likelihood Noise Suppression Filter," Tech. Note 1979-31, MIT Lincoln Lab., Lexington, Ma., June 1979.
9. J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," Proc. IEEE, *67*, No. 12 (December 1979), pp. 1586–604.
10. I. B. Thomas and R. J. Niederjohn, "Enhancement of Speech Intelligibility at High Noise Levels by Filtering and Clipping," J. Aud. Eng. Soc., *16*, No. 10 (October 1968), pp. 412–15.
11. R. A. Gillman, "A Fast Frequency Domain Pitch Algorithm," J. Acoust. Soc. Am., *58*, Supplement No. 1 (Fall 1975), p. S62.