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., narrow-band 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.

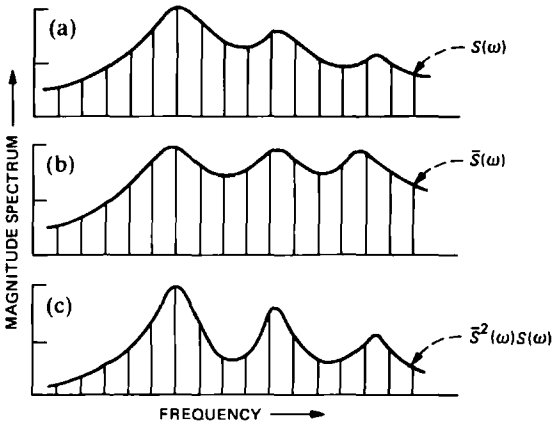

Fig. 1—Illustration of noise stripping by increasing the dynamic range between formant peaks and noise valleys. (a) Original spectral envelope and fine structure. (b) Formant-level equalized spectral envelope. (c) The product spectrum $\bar{S}^2(\omega)S(\omega)$ in which the ratios between formant peaks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

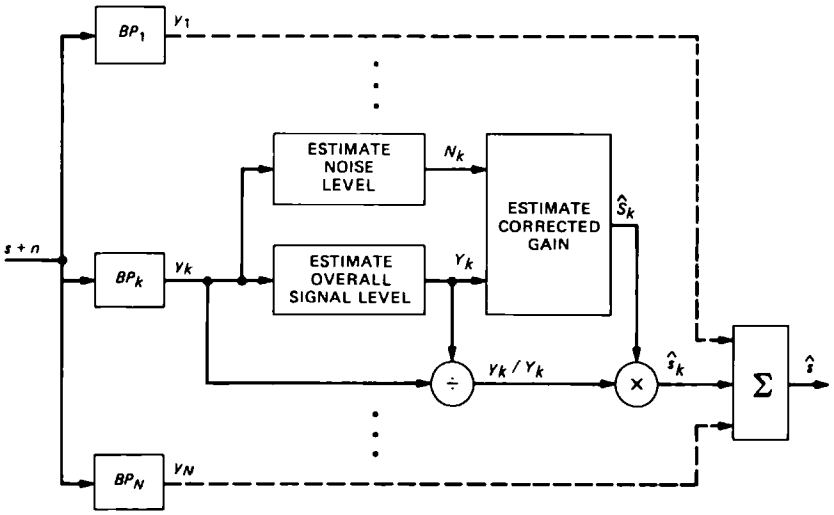

Fig. 2—Block diagram of the self-excited channel bank noise stripper consisting of a bank of N FIR bandpass filters with gain estimation and correction within each channel.

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

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.