Enhancement and Bandwidth Compression of Noisy Speech

JAE S. LIM, MEMBER, IEEE, AND ALAN V. OPPENHEIM, FELLOW, IEEE

Invited Paper

Abstract—Over the past several years there has been considerable attention focused on the problem of enhancement and bandwidth compression of speech degraded by additive background noise. This interest is motivated by several factors including a broad set of important applications, the apparent lack of robustness in current speech-compression systems and the development of several potentially promising and practical solutions. One objective of this paper is to provide an overview of the variety of techniques that have been proposed for enhancement and bandwidth compression of speech degraded by additive background noise. A second objective is to suggest a unifying framework in terms of which the relationships between these systems is more visible and which hopefully provides a structure which will suggest fruitful directions for further research.

I. INTRODUCTION

THERE ARE a wide variety of contexts in which it is desired to enhance speech. The objective of enhancement may perhaps be to improve the overall quality, to increase intelligibility, to reduce listener fatigue, etc. Depending on the specific application, the enhancement system may be directed at only one of these objectives or several. For example, a speech communication system may introduce a low-amplitude long-time delay echo or a narrow-band additive disturbance. While these degradations may not by themselves reduce intelligibility for the purposes for which the channel is used, they are generally objectionable and an improvement in quality perhaps even at the expense of some intelligibility may be desirable. Another example is the communication between a pilot and an air traffic control tower. In this environment, the speech is typically degraded by background noise. Of central importance is the intelligibility of the speech and it would generally be acceptable to sacrifice quality if the intelligibility could be improved. Even with normal undegraded speech, it is sometimes useful or desirable to provide enhancement. As a simple example high-pass filtering of normal speech is often used to introduce a "crispness" which is generally perceived as an improvement in quality.

The speech-enhancement problem covers a broad spectrum of constraints, applications and issues. Environments in which an additive background signal has been introduced are common. The background may be noise-like such as in aircraft, street noise, etc. or may be speech-like such as an environment with competing speakers. Other examples in which the need for speech enhancement arises include correcting for reverberation, correcting for the distortion of the speech of underwater divers breathing a helium-oxygen mixture, and correcting the distortion of speech due to pathological difficulties of the speaker or introduced due to an attempt to speak too rapidly. Even for these examples, the problem and techniques vary, depending on the availability of other signals or information. For example, for enhancement of speech in an aircraft a separate microphone can be used to monitor the background noise so that the characteristics of the noise can be used to adjust or adapt the enhancement system. At the air-traffic control tower, however, the only signal available for enhancement is the degraded speech.

Another very important application for speech enhancement is in conjunction with speech bandwidth compression systems. Because of the increasing role of digital communication channels coupled with the need for encrypting of speech and increased emphasis on integrated voice-data networks, speech-bandwidth-compression systems are destined to play an increasingly important role in speech-communication systems. The conceptual basis for narrow-band speech-compression systems stems from a model for the speech signal based on what is known about the physics and physiology of speech production. Because of this reliance on a model for the signal it is not unreasonable to expect that as the signal deviates from the model due to distortion such as additive noise, the performance of the speech compression system with regard to factors such as quality, intelligibility, etc., will degrade. It is generally agreed that the performance of current speech-compression systems degrades rapidly in the presence of additive noise and other distortions and there is currently considerable interest and attention being directed at the development of more robust speech compression systems. There are two basic approaches which are typically considered either of which may be preferable in a given situation. One approach is to base the bandwidth compression on the assumption of undistorted speech and develop a preprocessor to enhance the degraded speech in preparation for further processing by the bandwidth compression system. It is important to recognize that in enhancing speech in preparation for bandwidth compression the effectiveness of the preprocessor is judged on the basis of the output of the bandwidth-compression system in comparison with the output if no preprocessor is used. Thus, for example, it is possible that the output of the preprocessor would be judged by a listener to be inferior (by some measure) to the input but that the output of the bandwidth-compression system with the preprocessor is preferred to the output without it. In this case, the preprocessor would clearly be considered to be effective

Manuscript received June 22, 1979; revised August 28, 1979. This work was supported in part by the Defense Advance Research Projects Agency monitored by the Office of Naval Research under Contract N00014-75-C-0951-NR049-328 at M.I.T. Research Laboratory of Electronics and in part by the Department of the Air Force under Contract F19628-78-C-0002 at M.I.T. Lincoln Laboratory.

The authors are with M.I.T. Research Laboratory of Electronics and M.I.T. Lincoln Laboratory, Cambridge, MA 02139.

0018-9219/79/1200-1586$00.75

in enhancing the speech in preparation for bandwidth compression. Another approach to bandwidth compression of degraded speech is to incorporate into the model for the signal information about the degradation. A number of systems based on such an approach have recently been proposed and will be discussed in detail in this paper.

As is evident from the above discussion, the general problem of enhancing speech is broad and the constraints, information, and objectives are heavily dependent on the specific context and applications. In this paper, we consider only a small subset of possible topics, specifically the enhancement and bandwidth compression of speech degraded by additive noise. Furthermore, we assume that the only signal available is the degraded speech and that the noise does not depend on the original speech. Many practical problems, some of which have already been discussed, fall into this framework and some problems that do not can be transformed so that they do. For example, multiplicative noise or convolutional noise degradation can be converted to an additive noise degradation by a homomorphic transformation [1], [2]. As another example, signal-dependent quantization noise in pulse-code modulation (PCM) signal coding can be converted to a signal independent additive noise by a pseudo-noise technique [3]-[5].

Even within the limited framework outlined above, there is a diversity of approaches and systems. One objective of this paper is to provide an overview of the variety of techniques that have been proposed for enhancement of speech degraded by additive background noise both for direct listening and as a preprocessor for subsequent bandwidth compression. Many of these systems were developed independently of each other and on the surface often appear to be unrelated. Thus another objective of the paper is to provide a unifying framework in terms of which the relationship between these systems is more visible, and which hopefully will provide a structure which will suggest further fruitful directions for research.

In Section II, we present an overview of the general topic. In this overview we classify the various enhancement systems based on the information assumed about the speech and the noise. Some systems based on time-invariant Wiener filtering, for example, rely only on an assumed noise power spectrum and on long-time average characteristics of speech, such as the fact that the average speech spectrum decays with frequency at approximately 6 dB/octave. Other systems rely on aspects of speech perception or speech production in general or on a detailed model of speech.

Sections III-V present a more detailed discussion of several of these categories of speech-enhancement systems. In particular, Section III is concerned with the general principle of speech enhancement based on estimation of the short-time spectral amplitude of the speech. This basic principle encompasses a variety of techniques and systems including the specific methods of spectral subtraction, parametric Wiener filtering, etc. In Section IV, speech enhancement techniques which rely principally on the concept of the short-time periodicity of voiced speech are reviewed, including comb-filtering and related systems. Section V discusses a variety of systems that rely on more specific modeling of the speech waveform. As we will discuss in detail, in some cases, parameters of the model are obtained from an analysis of the degraded speech and used to synthesize the enhanced speech. In other cases, the results of an analysis based on a model for speech are used to control an enhancement filter, perhaps with the procedure being iterative so that the output of an enhancement filter is then subjected to further analysis, etc. Many of these systems also incorporate a number of the techniques introduced in Section III, including Wiener filtering and spectral subtraction.

In Sections III-V, the focus is entirely on systems for enhancement with the evaluation of the systems being based on listening without further processing. In Section VI, we consider the related but separate problem of bandwidth compression of speech degraded by additive noise.

In Section VII, we discuss in some detail the evaluation of the performance of the various systems presented in the earlier sections. In general, the performance evaluation of a speech-enhancement system is extremely difficult, in large measure because the appropriate criteria for evaluation are heavily dependent on the specific application of the system. Relative importance of such factors as quality, intelligibility, listener fatigue, etc., may vary considerably with the application. In Section VII, we summarize the performance evaluations that have been reported for the various systems presented in this paper. Since the evaluation of different systems has generally been based on different procedures, environments, etc., no attempt is made in the section to *compare* individual systems. In general, however, we will see that while many of the enhancement systems reduce the apparent background noise and thus perhaps increase quality, many of them to varying degrees, reduce intelligibility. In the context of bandwidth compression, however, various systems provide an increase in intelligibility over that obtained without the incorporation of speech enhancement.

II. Overview of Systems for Enhancement and Bandwidth Compression of Noisy Speech

As indicated in the previous section, our focus in this paper is on degradation due to the presence of additive noise. Even within this limited context there are a wide variety of approaches which have been proposed and explored. Conceptually any approach should attempt to capitalize on available information about the signal, i.e., the speech, and the background noise. Speech is a special subclass of audio signals and there are reasonable models in terms of which the speech waveform can be described and categorized. The more specifically we attempt to model the speech signal, the more potential for separating it from the background noise. On the other hand, the more we assume about the speech the more sensitive the enhancement system will be to inaccuracies or deviations from these assumptions. Thus incorporating assumptions and information about the speech signal represents tradeoffs which are reflected in the various systems. In a similar manner systems can attempt to incorporate detailed information about the background noise. For example, the type of processing suggested if the background noise is a competing speaker is different than if it is wide-band random noise. Thus enhancement systems also tend to differ in terms of the assumptions made regarding the background noise. As with assumptions related to the signal, the more an enhancement system attempts to capitalize on assumed characteristics of the noise the more susceptible it is likely to be to deviations from these assumptions.

Another important consideration in speech enhancement stems from the fact that the criteria for enhancement ultimately relate to an evaluation by a human listener. In different contexts the criteria for evaluation may differ depending on whether quality, intelligibility, or some other attribute is the

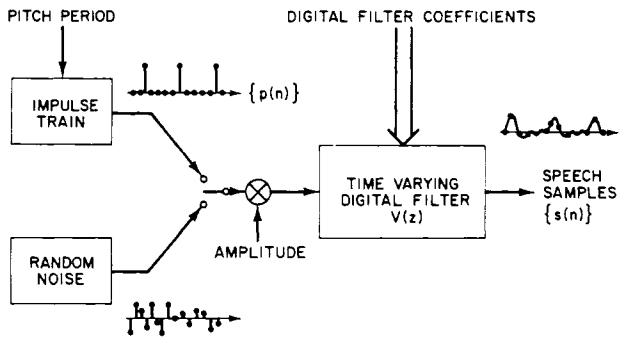

Fig. 1. A speech production model.

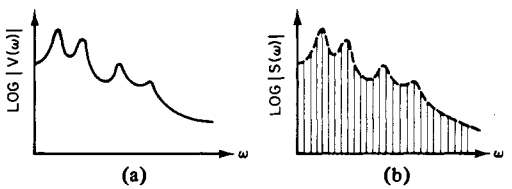

Fig. 2. An example of resonant frequencies of an acoustic cavity. (a) Vocal-tract transfer function. (b) Magnitude spectrum of a speech sound with the resonant frequencies shown in (a).

most important. Thus speech enhancement must inevitably take into account aspects of human perception. As we will indicate shortly, some systems are heavily motivated by perceptual considerations, others rely more on mathematical criteria. In such cases, of course, the mathematical criteria must in some way be consistent with human perception, and, while an optimum mathematical criterion is not known, some mathematical error criteria are understood to be a better match than others to aspects of human perception.

In the following discussion we briefly describe some aspects of speech production and speech perception that in varying degrees play a role in speech-enhancement systems. Following that we present a brief overview of a representative collection of speech-enchancement systems, with the intent of categorizing these systems in terms of the various aspects of speech production and perception on which they attempt to capitalize.

Speech is generated by exciting an acoustic cavity, the vocal tract, by pulses of air released through the vocal cords for voiced sounds, or by turbulence for unvoiced sounds. Thus a simple but useful model for speech production consists of a linear system, representing the vocal tract, driven by an excitation function which is a periodic pulse train for voiced sounds and wide-band noise for unvoiced sounds, as illustrated in Fig. 1. Furthermore, since the linear system represents an acoustic cavity, its response is of a resonant nature, so that its transfer function is characterized by a set of resonant frequencies, referred to as formants, as illustrated in Fig. 2(a). Thus, if the excitation and vocal-tract parameters are fixed, then as indicated in Fig. 2(b), the speech spectrum has an envelope representing the vocal-tract transfer function of Fig. 2(a) and a fine structure representing the excitation.

Many of the techniques for speech enhancement, particularly those in Sections III and V are conceptually based on the representation of the speech signal as a stochastic process. This characterization of speech is clearly more appropriate in the case of unvoiced sounds for which the vocal tract is driven by wide-band noise. The vocal tract of course changes shape as different sounds are generated and this is reflected in a time varying transfer function for the linear system in Fig. 1. However, because of the mechanical and physiological constraints on the motion of the vocal tract and articulators such as the tongue and lips, it is reasonable to represent the linear system in Fig. 1 as a slowly varying linear system so that on a short-time basis it is approximated as stationary. Thus some specific attributes of the speech signal, which can be capitalized on in an enhancement system are that it is the response of a slowly varying linear system, that on a short-time basis its spectral envelope is characterized by a set of resonances, and that for voiced sounds, on a short-time basis it has a harmonic structure. This simplified model for speech production has generally been very successful in a variety of engineering contexts including speech enchancement, synthesis, and bandwidth compression. A more detailed discussion of models for speech production can be found in [6]-[8].

The perceptual aspects of speech are considerably more complicated and less well understood. However, there are a number of commonly accepted aspects of speech perception which play an important role in speech-enchancement systems. For example, consonants are known to be important in the intelligibility of speech even though they represent a relatively small fraction of the signal energy. Furthermore, it is generally understood that the short-time spectrum is of central importance in the perception of speech and that, specifically, the formants in the short-time spectrum are more important than other details of the spectral envelope. It appears also, that the first formant, typically in the range of 250 to 800 Hz, is less important perceptually, than the second formant [9], [10]. Thus it is possible to apply a certain degree of high pass filtering [11], [12] to speech which may perhaps affect the first formant without introducing serious degradation in intelligibility. Similarly low-pass filtering with a cutoff frequency above 4 kHz, while perhaps affecting crispness and quality will in general not seriously affect intelligibility. A good representation of the magnitude of the short-time spectrum is also generally considered to be important whereas the phase is relatively unimportant. Another perceptual aspect of the auditory system that plays a role in speech enhancement is the ability to mask one signal with another. Thus, for example, narrow-band noise and many forms of artificial noise or degradation such as might be produced by a vocoder are more unpleasant to listen to than broad-band noise and a speech-enhancement system might include the introduction of broad-band noise to mask the narrow-band or artificial noise.

All speech-enhancement systems rely to varying degrees on the aspects of speech production and perception outlined above. One of the simplest approaches to enhancement is the use of low-pass or bandpass filtering to attenuate the noise outside the band of perceptual importance for speech. More generally, when the power spectrum of the noise is known, one can consider the use of Wiener filtering, based on the long-time power spectrum of speech. While in some cases such as the presence of narrow-band background noise, this is reasonably successful, Wiener filtering based on the long-time power spectrum of the speech and noise is limited because speech is not stationary. Even if speech were truly stationary, mean-square error which is the error criterion on which Wiener filtering is based is not strongly correlated with perception and thus is not a particularly effective error criterion to apply to speech processing systems. This is evidenced, for example, in the use of masking for enhancement. By adding broad-band

noise to mask other degradation, we are, in effect, increasing the mean-square error. Another example that suggests that mean-square error is not well matched to the perceptually important attributes in speech is the fact that distortion of the speech waveform by processing with an all-pass filter results in essentially no audible difference if the impulse response of the all-pass filter is reasonably short but can result in a substantial mean-square error between the original and filtered speech. In other words, mean-square error is sensitive to phase of the spectrum whereas perception tends not to be.

Masking and bandpass filtering represent two simple ways in which perceptual aspects of the auditory system can be exploited in speech enhancement. Another system whose motivation depends heavily on aspects of speech perception was proposed by Thomas and Niederjohn [12] as a preprocessor prior to the introduction of noise in those applications where noise-free speech is available for processing. In essence, their system applies high-pass filtering to reduce or remove the first formant followed by infinite clipping. The motivation for the system lies in the observation that at a given signal-to-noise ratio infinite clipping will increase, relative to the vowels, the amplitude of the perceptually important low-amplitude events such as consonants thus making them less susceptible to masking by noise. In addition, for vowels the filtering will increase the amplitude of higher formants relative to the first formant, thus making the perceptually more important higher formants less susceptible to degradation. In the speech enhancement problem considered in this paper, noise-free speech is not available for processing as required in the above system. Thomas and Ravindran [13], however, applied high-pass filtering followed by infinite clipping to noisy speech as an experiment. While quality may be degraded by the process of filtering and clipping, they claim a noticeable improvement in intelligibility when applied to enhance speech degraded by wide-band random noise. One possible explanation may be that the high-pass filtering operation reduces the masking of perceptually important higher formants by the relatively unimportant low-frequency components.

Another system which relies heavily on human perception of speech was proposed by Drucker [14]. Based on some perceptual tests, Drucker concluded that one primary cause for the intelligibility loss in speech degraded by wide-band random noise is the confusion among the fricative and plosive sounds which is partly due to the loss of short pauses immediately before the plosive sounds. By high-pass filtering one of the fricative sounds, the /s/ sound, and inserting short pauses before the plosive sounds (assuming that their locations can be accurately determined), Drucker claims a significant improvement in intelligibility.

In discussing perceptual attributes we indicated that the short-time spectral magnitude is generally considered to be important whereas the phase is relatively unimportant. This forms the basis for a class of speech enhancement systems which attempt in various ways to estimate the short-time spectral magnitude of the speech without particular regard to the phase and to use this to recover or reconstruct the speech. This class of systems includes spectral subtraction techniques originally due to Weiss *et al.* [15], [16], and which have recently received a great deal of attention [17]–[22] and optimum filtering techniques such as Wiener filtering and power spectrum filtering. These systems will be discussed in considerable detail in Section III. As we will see, many of these systems which appear on the surface to be different are in fact identical or very closely related.

In addition to directly or indirectly utilizing perceptual attributes most enhancement systems rely to varying degrees on aspects of speech production. For example, in Section IV, we describe in detail a variety of systems that attempt, in some way, to capitalize on short-time periodicity of speech during voiced sounds. As a consequence of this periodicity, during voiced intervals the speech spectrum has a harmonic structure which suggests the possibility of applying comb filtering or as proposed by Parsons [23] attempting to extract in other ways, the components of the speech spectrum only at the harmonic frequencies. In essence, knowledge of the harmonic structure of voiced sounds allows us in principle to remove the noise in the spectral bands between the harmonics.

As discussed in Section IV, speech enhancement by comb filtering can also be viewed in terms of averaging successive periods of the noisy speech to partially cancel the noise. Another system, which attempts to take advantage of the quasi-periodic nature of the speech was proposed by Sambur [24]. As developed in more detail in Section IV, his system is based on the principles of adaptive noise cancelling. Unlike the classical procedure Sambur's method is designed to cancel out the clean speech signal, taking advantage of the quasi-periodic nature of the speech to form an estimate of the speech at each time instant from the value of the signal one period earlier.

In the model of speech production, we represented the speech signal as generated by exciting a quasi-stationary linear system with a pulse train for voiced speech and noise for unvoiced speech. Based on this model, an approach to speech enhancement is to attempt to estimate parameters of the model rather than the speech itself and to then use this to synthesize the speech, i.e., to enhance speech through the use of an analysis–synthesis system. A particularly novel application of this concept was used by Miller [25] to remove the orchestral accompaniment from early recordings of Enrico Caruso. In this system homomorphic deconvolution was used to estimate the impulse response of the model in Fig. 1. A similar approach to noise reduction was proposed by Suzuki [26], [27] whereby the short-time correlation function of the degraded speech is used as an estimate of the impulse response of the linear system. This system is referred to as splicing of auto correlation function (SPAC). A modification of SPAC is referred to as splicing of cross-correlation function (SPOC). A number of systems also attempt to model the vocal-tract impulse response in more detail. As we discussed previously the vocal-tract transfer function is characterized by a set of resonances or formants that are perceptually important. This suggests the possibility of representing the vocal-tract impulse response in terms of a pole-zero model with the analysis procedure directed at estimating the associated parameters. The poles in particular would provide a reasonable representation of the formants.

All-pole modeling of speech has had notable success in analysis–synthesis systems for clean speech. A number of recent efforts have been directed toward estimating the parameters in an all-pole model from noisy observations of the speech such as the systems by Magill and Un [28], Lim and Oppenheim [29], Lim [18], and Done and Rushforth [30]. Extensions to pole-zero modeling have also been proposed

by Musicus and Lim [31] and Musicus [32]. These various approaches are described and compared in detail in Section V.

The above discussion was intended as a brief overview of the general approaches to speech enhancement. In the next three sections we explore in more detail many of the systems mentioned above. In particular, in Section III, we focus on speech-enhancement techniques based on short-time spectral amplitude estimation. In Section IV our focus is on speech enhancement based on periodicity of voiced speech and in Section V on speech-enhancement techniques using an analysis–synthesis procedure.

III. Speech Enhancement Techniques Based on Short-Time Spectral Amplitude Estimation

In general, in enhancement of a signal degraded by additive noise, it is significantly easier to estimate the spectral amplitude associated with the original signal than it is to estimate both amplitude and phase. As we discussed in Section II, it is principally the short-time spectral amplitude rather than phase that is important for speech intelligibility and quality. As we discuss in this section, there are a variety of speech-enhancement techniques that capitalize on this aspect of speech perception by focusing on enhancing only the short-time spectral amplitude. The techniques to be discussed can be broadly classified into two groups. In the first, presented in Section III-A, the short-time spectral amplitude is estimated in the frequency domain, using the spectrum of the degraded speech. Each short-time segment of the enhanced speech waveform in the time domain is then obtained by inverse transforming this spectral amplitude estimate combined with the phase of the degraded speech. In the second class, discussed in Section III-B the degraded speech is first used to obtain a filter which is then applied to the degraded speech. Since these procedures lead to zero-phase filters, it is again only the spectral amplitude that is enhanced, with the phase of the filtered speech being identical to that of the degraded speech.

In both classes of systems discussed below no conceptual distinction is made between voiced and unvoiced speech and in particular in contrast to the techniques to be discussed in Section IV the periodicity of voiced speech is not exploited. Both classes of systems in this section are most easily interpreted in terms of a stochastic characterization of the speech signal. While this characterization is more justifiable for unvoiced speech it has been shown empirically to also lead to successful procedures for voiced speech.

A. Speech Enhancement Based on Direct Estimation of Short-Time Spectral Amplitude

When a stationary random signal $s(n)$ has been degraded by uncorrelated additive noise $d(n)$ with a known power density spectrum, the power density spectrum or spectral amplitude of the signal is easily estimated through a process of spectral subtraction. Specifically, if

$$y(n) = s(n) + d(n) \tag{1}$$

and $P_y(\omega)$, $P_s(\omega)$, and $P_d(\omega)$ represent the power density spectra of $y(n)$, $s(n)$, and $d(n)$, respectively, then

$$P_y(\omega) = P_s(\omega) + P_d(\omega). \tag{2}$$

Consequently, a reasonable estimate for $P_s(\omega)$ is obtained by subtracting the known spectrum $P_d(\omega)$ from an estimate of $P_y(\omega)$ developed from the observations of $y(n)$.

Speech, of course, is not a stationary signal. However, with $s(n)$ in (1) now representing a speech signal and with the processing to be carried out on a short-time basis we consider $s(n)$, $d(n)$, and $y(n)$ multiplied by a time-limited window $w(n)$. With $y_w(n)$, $d_w(n)$, and $s_w(n)$ denoting the windowed signals $y(n)$, $d(n)$, and $s(n)$ and $Y_w(\omega)$, $D_w(\omega)$, and $S_w(\omega)$ as their respective Fourier transforms we have

$$y_w(n) = s_w(n) + d_w(n) \tag{3}$$

and

$$|Y_w(\omega)|^2 = |S_w(\omega)|^2 + |D_w(\omega)|^2 + S_w(\omega) \cdot D_w^*(\omega) + S_w^*(\omega) \cdot D_w(\omega) \tag{4}$$

where $D_w^*(\omega)$ and $S_w^*(\omega)$ represent complex conjugates of $D_w(\omega)$ and $S_w(\omega)$. The function $|S_w(\omega)|^2$ will be referred to as the short-time energy spectrum of speech. For speech enhancement based on the short-time spectral amplitude, the objective is to obtain an estimate $|\hat{S}_w(\omega)|$ of $|S_w(\omega)|$ and from this, an estimate $\hat{s}_w(n)$ of $s_w(n)$.

From the estimate $\hat{s}_w(n)$, speech can be generated in a variety of different ways. One approach is to use an analysis window function $w(n)$ that generates $s(n)$ when all the frames of $s_w(n)$ are overlapped and added with the appropriate time registration. Such a window function satisfies the equation

$$\sum_i w_i(n) = 1, \quad \text{for all } n \text{ of interest} \tag{5}$$

where $w_i(n)$ represents the ith window frame. Two such examples are overlapped triangular and hamming windows. Using such a window function, speech is then generated by adding up the estimates of the windowed segments.

Various speech-enhancement techniques discussed in this section differ primarily in how $|S_w(\omega)|$ is specifically estimated from the noisy speech. In one spectral subtraction technique referred to as *power spectrum subtraction*,[1] $|S_w(\omega)|$ is estimated based on (4). From the observed data $y_w(n)$, $|Y_w(\omega)|^2$ can be obtained directly. The terms $|D_w(\omega)|^2$, $S_w(\omega) \cdot D_w^*(\omega)$ and $S_w^*(\omega) \cdot D_w(\omega)$ cannot be obtained exactly and in the power spectrum subtraction technique they are approximated by $E[|D_w(\omega)|^2]$, $E[S_w(\omega) \cdot D_w^*(\omega)]$ and $E[S_w^*(\omega) \cdot D_w(\omega)]$ where $E[\cdot]$ denotes the ensemble average. For $d(n)$ zero mean[2] and uncorrelated with $s(n)$, $E[S_w(\omega) \cdot D_w^*(\omega)]$ and $E[S_w^*(\omega) \cdot D_w(\omega)]$ are zero and an estimate $|\hat{S}_w(\omega)|^2$ of $|S_w(\omega)|^2$, is suggested from (4) as

$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - E[|D_w(\omega)|^2], \tag{6}$$

where $E[|D_w(\omega)|^2]$ is obtained either from the assumed known properties of $d(n)$ or by an actual measurement from the background noise in the intervals where speech is not present. The estimate $|\hat{S}_w(\omega)|^2$ based on (6) is not guaranteed to be nonnegative since the right-hand side can become negative, and a number of somewhat arbitrary choices have been made. In some studies, the negative values are made positive by changing the sign. In some other studies $|\hat{S}_w(\omega)|^2$ is set to zero if

[1] The name "power spectrum subtraction" comes from the close similarity between (2) and (6).

[2] The zero mean assumption for the additive random noise is made only for notational convenience.

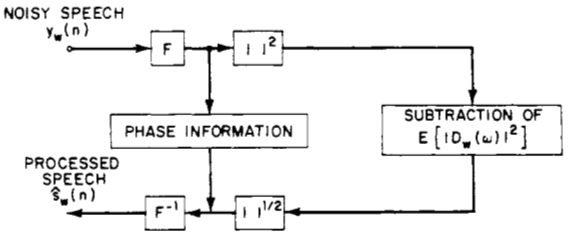

Fig. 3. A typical speech enhancement system based on the power spectrum (correlation) subtraction technique. The negative result after subtraction is set to zero.

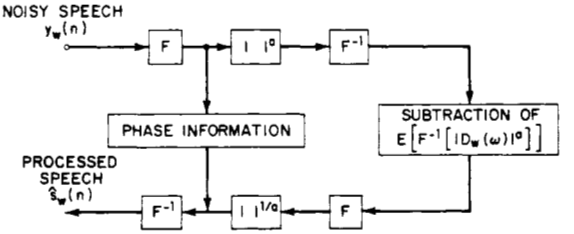

Fig. 4. A speech-enhancement system proposed by Weiss *et al.* [15], [16].

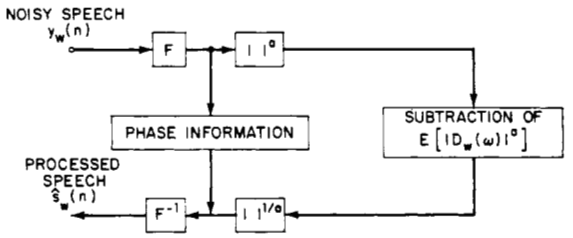

Fig. 5. A simplification of the system in Fig. 4. The negative result after subtraction is set to zero.

$|Y_w(\omega)|^2$ is less than $E[|D_w(\omega)|^2]$. The latter approach has been more extensively used in the literature, and as will be seen later it can be related directly to the optimum filtering technique discussed in Section III-B.

Given an estimate of $|S_w(\omega)|$, there are a variety of different ways to estimate $s_w(n)$. One method which has been used extensively in the class of systems discussed in this section and is also consistent with the notion that short-time phase is relatively unimportant is to approximate $\measuredangle S_w(\omega)$, the phase of $S_w(\omega)$, by $\measuredangle Y_w(\omega)$, so that

$$\hat{S}_w(\omega) = |\hat{S}_w(\omega)| \cdot \exp[j \measuredangle Y_w(\omega)] \tag{7}$$

and

$$\hat{s}_w(n) = F^{-1}[\hat{S}_w(\omega)]. \tag{8}$$

A typical algorithm for speech enhancement by the power spectrum subtraction technique is shown in Fig. 3.

Except for some details and interpretations, the power spectrum subtraction technique discussed above is a special case of a more general system originated by Weiss *et al.* [15], [16]. Specifically, the power spectrum subtraction technique can also be interpreted in terms of estimating the short-time correlation $\phi_s(n)$ as

$$\phi_s(n) = \phi_y(n) - E[\phi_d(n)] \tag{9}$$

where

$$\phi_s(n) = \sum_{k=-\infty}^{\infty} s_w(k) \cdot s_w(k-n) = F^{-1}[|S_w(\omega)|^2] \tag{10}$$

and $\phi_y(n)$ and $\phi_d(n)$ are similarly defined. For this reason, the power spectrum subtraction technique is also referred to as the *correlation subtraction* technique. Weiss *et al.* focused on estimating the short-time correlation function and in place of a squaring operation used an arbitrary positive real constant "a". In their technique, then, by defining $\tilde{\phi}_s(n)$ to be $F^{-1}[|S_w(\omega)|^a]$, $\tilde{\phi}_s(n)$ is estimated as

$$\begin{aligned}\hat{\tilde{\phi}}_s(n) &= \tilde{\phi}_y(n) - E[\tilde{\phi}_d(n)] \\ &= F^{-1}[|Y_w(\omega)|^a] - E[F^{-1}[|D_w(\omega)|^a]].\end{aligned} \tag{11}$$

Based on this estimate and the assumption that $\measuredangle S_w(\omega)$ equals $\measuredangle Y_w(\omega)$, the windowed speech $s_w(n)$ is estimated. The speech enhancement system proposed by Weiss *et al.* is shown in Fig. 4.

The system in Fig. 4 can be simplified both computationally and conceptually [18], [19] by recognizing that the expectation and Fourier transform operations in (11) are interchangeable and therefore (11) is equivalent to

$$|\hat{S}_w(\omega)|^a = |Y_w(\omega)|^a - E[|D_w(\omega)|^a]. \tag{12}$$

Such a simplified system based on (12) is shown in Fig. 5. As is evident in Fig. 5 the system proposed by Weiss *et al.* is a technique to estimate the short-time spectral amplitude of speech by a particular form of spectral subtraction. The performance of the system in Fig. 5 as a speech enhancement system was evaluated by Lim (19) and the results will be discussed in Section VII. When the constant "a" is set to unity, the system in Fig. 5 reduces to the speech enhancement system developed by Boll [20].

The parameter "a" in [12] obviously affords a degree of flexibility over the system based on (6). A further generalization is to introduce an additional degree of flexibility by estimating $|S_w(\omega)|$ through the relation

$$|\hat{S}_w(\omega)|^a = |Y_w(\omega)|^a - kE[|D_w(\omega)|^a] \tag{13}$$

where now there are the two parameters a and k. This generalization with a and k as parameters was considered for speech enhancement by Lim [18] and Berouti *et al.* [21]. Just as with the specific form of spectral subtraction in (6), each short-time speech segment is in effect estimated by utilizing the phase associated with the noisy speech, and negative values on the right-hand side of (13) can be dealt with through the use of full-wave or half-wave rectification. The additional possibility of also utilizing a frequency dependent threshold on the right-hand side of (13) was considered by Berouti *et al.* [21].

Another approach, which leads to a further modification of spectral subtraction was proposed by McAulay and Malpass [33]. In this approach, the problem was formulated by assuming that at each frequency the noise is Gaussian and developing the maximum likelihood estimate of $|S_w(\omega)|$. The resulting estimate has the form

$$|\hat{S}_w(\omega)| = \tfrac{1}{2}|Y_w(\omega)| + \tfrac{1}{2}[|Y_w(\omega)|^2 - E[|D_w(\omega)|^2]]^{1/2}. \tag{14}$$

A further variation, proposed by McAulay and Malpass [33] modifies (14) by a factor which is chosen to represent as a function of $|Y_w(\omega)|$ the probability that speech is in fact present in the signal $y(n)$. Modification of (14) by this factor is based on the notion that as the probability that only noise is present increases, it might perhaps be preferable to further

reduce the estimate of $|S_w(\omega)|$. Other techniques for speech enhancement similar or very closely related to the various spectral subtraction techniques discussed above include the work of Curtis and Niederjohn [17] and Preuss [22].

In this section, we have discussed a variety of different techniques to estimate the short time spectral amplitude of speech. Many of them can be viewed as attempting to enhance the speech-to-noise (S/N) ratio by not affecting the spectral components corresponding to relatively high S/N ratio but attenuating those corresponding to relatively low S/N ratio. To illustrate this point, consider the spectral subtraction method corresponding to (13) with the assumption that $a = 2$ and that the right-hand side is positive. Expressing the estimate $\hat{S}_w(\omega)$ in the form of a zero-phase frequency response $H(\omega)$ applied to $Y_w(\omega)$,

$$H(\omega) = \frac{Y_w(\omega)}{\hat{S}_w(\omega)} = \left(\frac{|Y_w(\omega)|^2 - k \cdot E[|D_w(\omega)|^2]}{|Y_w(\omega)|^2} \right)^{1/2}. \quad (15)$$

Equation (15) can be rewritten as

$$H(\omega) = \left(\frac{X^2(\omega) - k}{X^2(\omega)} \right)^{1/2} \quad (16)$$

where

$$X^2(\omega) = \frac{|Y_w(\omega)|^2}{E[|D_w(\omega)|^2]}. \quad (17)$$

From (17), $X(\omega)$ can be interpreted as a speech-plus-noise-to-noise ratio at each frequency ω. In Fig. 6 is plotted 20 log $H(\omega)$ for different values of the constant "k" as a function of 20 log $X(\omega)$. It is clear from the figure that the frequency components of $Y_w(\omega)$ corresponding to low S/N ratio are severely attenuated. As another example, a similar plot representing the speech enhancement system corresponding to (14) derived from maximum likelihood considerations (33) is also shown in Fig. 6. The results in Fig. 6 are generally applicable to various short-time spectral amplitude estimation techniques discussed in this section and will be useful in understanding the results of the performance evaluation discussed in Section VII.

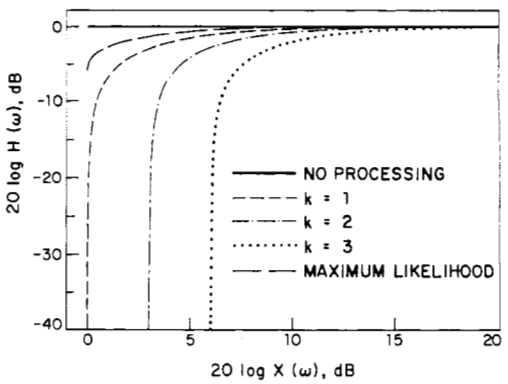

Fig. 6. Attenuation curves for two spectral subtraction techniques (equations (13) and (14)). See text for details.

B. Speech Enhancement Techniques Based on Wiener Filtering

In the previous section, the basis for enhancement was the explicit estimation of the short-time spectral magnitude through a process of spectral subtraction. In this section, we discuss techniques in which a frequency weighting for an "optimum" filter is first estimated from the noisy speech. This filter is then applied either in the time domain or frequency domain to obtain an estimate of the undegraded speech. Thus with $Y_w(\omega)$, $D_w(\omega)$, and $S_w(\omega)$ again denoting the short-time spectra associated with the windowed time functions $y(n)$, $d(n)$, and $s(n)$, the estimate $\hat{S}_w(\omega)$ of $S_w(\omega)$ takes the form

$$\hat{S}_w(\omega) = H(\omega) Y_w(\omega). \quad (18)$$

As we saw in (15), the techniques in Section III-A can also be put into this form and consequently the essential difference between the techniques presented in that section and those to be discussed here rests in the basis on which the frequency weighting $H(\omega)$ is obtained. In this section we focus on procedures for obtaining $H(\omega)$ based on the principles of Wiener filtering. However, as we will see toward the end of this section, an implicit form of this procedure leads, in fact to frequency weightings identical to several discussed in Section III-A.

As is well known, for $y(n) = s(n) + d(n)$ in which $s(n)$ and $d(n)$ represent uncorrelated stationary random processes with power density spectra $P_s(\omega)$ and $P_d(\omega)$, respectively, the linear estimator of $s(n)$ which minimizes the mean-square error is obtained by filtering $y(n)$ with the noncausal Wiener filter for which the frequency response is

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)}. \quad (19)$$

The noncausal Wiener filter of (19) cannot be applied directly to estimate $s(n)$ since speech cannot be assumed to be stationary and the spectrum $P_s(\omega)$ cannot be assumed known. An approach often used is to approximate the noncausal Wiener filter with an adaptive Wiener filter with frequency response

$$H(\omega) = \frac{E[|S_w(\omega)|^2]}{E[|S_w(\omega)|^2] + E[|D_w(\omega)|^2]}. \quad (20)$$

As in Section III-A, the function $E[|D_w(\omega)|^2]$ may be obtained either from the assumed known statistics of $d(n)$ or by averaging many frames of $|D_w(\omega)|^2$ during silence intervals in which the statistics of the background noise can be assumed to be stationary. In estimating $E[|S_w(\omega)|^2]$, there are a variety of possibilities. Callahan [34] first estimates $E[|Y_w(\omega)|^2]$ by locally averaging $|Y_w(\omega)|^2$ over many frames of noisy speech. Then $E[|D_w(\omega)|^2]$ is subtracted from the estimated $E[|Y_w(\omega)|^2]$ to form an estimate of $E[|S_w(\omega)|^2]$. An equally reasonable method is to first estimate $E[|Y_w(\omega)|^2]$ by smoothing $|Y_w(\omega)|^2$ rather than averaging $|Y_w(\omega)|^2$ over many frames of noisy speech and then subtracting $E[|D_w(\omega)|^2]$ from the estimated $E[|Y_w(\omega)|^2]$. As other possibilities $E[|S_w(\omega)|^2]$ may be approximated as $|\hat{S}_w(\omega)|^2$ or by smoothing $|\hat{S}_w(\omega)|^2$ where $|\hat{S}_w(\omega)|^2$ is obtained from the short-time spectral amplitude estimation techniques discussed in Section III-A.

Given $H(\omega)$, the short-time speech segment is then obtained as specified by (18) applied either in the time domain or in the frequency domain. It should be noted that in all of the above procedures, the frequency weighting $H(\omega)$ has zero phase and thus from (18) the phase associated with the estimate $\hat{S}_w(\omega)$ is that of $Y_w(\omega)$. Thus just as with the procedures in Section III-A, it is only the spectral magnitude of $S_w(\omega)$ which is estimated.

Generalizations of Wiener filtering may also be considered. One such generalization which has been studied extensively [35], [36] in the context of image restoration has the fre-

quency response given by

$$H(\omega) = \left(\frac{P_s(\omega)}{P_s(\omega) + \alpha \cdot P_d(\omega)} \right)^{\beta} \tag{21}$$

for some constants "α" and "β" and has been referred to as parametric Wiener filters. By varying the constants "α" and "β", filters with different characteristics can be obtained. For example if α and β are unity, then (21) corresponds to Wiener filtering as specified in (19). If α is unity and β is 1/2, then (21) corresponds to power spectrum filtering (37) which has the characteristics that the enhanced signal has the same power spectrum $P_s(\omega)$ used in (21). Again, due to the nonstationarity of speech, equation (21) has to be modified. The approximation of $P_s(\omega)$ and $P_d(\omega)$ by the corresponding short-time energy spectra and generation of speech based on the estimated $H(\omega)$ have already been discussed. With this approximation, the frequency response associated with short-time parametric Wiener filtering would then be expressed as

$$H(\omega) = \left[\frac{E[|S_w(\omega)|^2]}{E[|S_w(\omega)|^2 + \alpha E[|D_w(\omega)|^2]} \right]^{\beta}. \tag{22}$$

In the Wiener filter of (20) or its generalized form of (22) it is assumed that the term representing $P_s(\omega)$ or $E[|S_w(\omega)|^2]$ is first obtained and the frequency weighting is then applied to $Y_w(\omega)$. An alternative is to treat (20) and (22) as implicit relationships. For example, let us estimate $E[|S_w(\omega)|^2]$ as

$$\hat{E}[|S_w(\omega)|^2] = |\hat{S}_w(\omega)|^2 \tag{23}$$

where $\hat{S}_w(\omega)$ is the estimate of the short-time spectrum of the speech. Then,

$$\hat{S}_w(\omega) = H(\omega) Y_w(\omega)$$

or

$$\hat{S}_w(\omega) = \left[\frac{|\hat{S}_w(\omega)|^2}{|\hat{S}_w(\omega)|^2 + \alpha E[|D_w(\omega)|^2]} \right]^{\beta} Y_w(\omega) \tag{24}$$

so that

$$|\hat{S}_w(\omega)| = \left[\frac{|\hat{S}_w(\omega)|^2}{|\hat{S}_w(\omega)|^2 + \alpha E[|D_w(\omega)|^2]} \right]^{\beta} |Y_w(\omega)|. \tag{25}$$

This, of course, is an implicit relationship, from which we would like to obtain $|\hat{S}_w(\omega)|$ and thus we refer to it as *implicit Wiener filtering*. For example, two solutions to (23) for $\beta = 1/2$ are

$$|\hat{S}_w(\omega)| = 0 \tag{26a}$$

$$|\hat{S}_w(\omega)| = [|Y_w(\omega)|^2 - \alpha E[|D_w(\omega)|^2]]^{1/2}. \tag{26b}$$

Thus, a solution for $|\hat{S}_w(\omega)|$ consistent with (25) is (26b) for positive values under the radical and (26a), i.e., zero otherwise. This, of course, is precisely the spectral subtraction method of (13) with $a = 2$. Similarly, for $\beta = 1$ a solution to (25) is

$$|\hat{S}_w(\omega)| = \tfrac{1}{2}|Y_w(\omega)| + \tfrac{1}{2}[|Y_w(\omega)|^2 - 4\alpha E[|D_w(\omega)|^2]]^{1/2} \tag{27}$$

For $\alpha = 1/4$ this is identical to the maximum likelihood estimate of (14).

Another potential generalization of Wiener filtering stems from considering an iterative approach to estimating

$$E[|S_w(\omega)|^2]$$

in (22). For example, let us consider an iterative procedure whereby $|\hat{S}_w(\omega)_i|$ denotes the estimate of $|S_w(\omega)|$ on the ith iteration with

$$S_w(\omega)_{i+1} = H_i(\omega) Y_w(\omega). \tag{28}$$

The transfer function $H_i(\omega)$ is in the form of (22) with $E[|S_w(\omega)|^2]$ estimated from $S_w(\omega)_i$. In such iterative procedures there are, of course, issues of convergence which will in general depend on the way in which the iteration is started and on specifically how $E[|S_w(\omega)|^2]$ is estimated from $S_w(\omega)_i$.

IV. Speech-Enhancement Techniques Based on Periodicity of Voiced Speech

In this section, we discuss speech enhancement techniques which capitalize on the observation that waveforms of voiced sounds are periodic with a period that corresponds to the fundamental frequency. Even with this basic underlying principle many different approaches are possible. In Section IV-A, we discuss an approach based on comb filtering to pass the harmonics of speech but reject the frequency components between the harmonics. In Section IV-B, we consider the extraction of speech harmonics from a high resolution spectrum of noisy speech. In Section IV-C, we discuss the use of adaptive noise cancelling techniques to reduce the background noise by capitalizing on the periodicity of voiced sounds to provide a reference input.

A. Speech Enhancement Based on Adaptive Comb Filtering

The periodicity of a time waveform manifests itself in the frequency domain as harmonics with the fundamental frequency corresponding to the period of the time waveform as is shown in Fig. 7. In Fig. 7(a) is shown a segment of a periodic time waveform and in Fig. 7(b) is shown the associated magnitude spectrum. Since the energy of a periodic signal is concentrated in bands of frequencies as is evident in Fig. 7(b) and the interfering signals in general have energy over the entire frequency bands, to the extent that accurate information of the fundamental frequency is available, a comb filter as shown in Fig. 7(c) can reduce noise while preserving the signal.

Even though voiced speech is only approximately periodic, the concept of comb filtering to reduce the background noise in noisy speech may still be applicable. One approach to enhancing degraded speech through comb filtering was taken by Shields [38]. A typical impulse response of a comb filter as applied by Shields is shown in Fig. 8(a). The spacing "T" in the figure represents the pitch period and a different value of "T" is chosen in processing different parts of voiced speech to adapt globally to the time varying nature of speech. Frazier *et al.* [39], observed that even with accurate fundamental frequency information Shields' adaptive comb filtering technique distorts speech signals significantly due to the time varying nature of speech even on a short-time (local) basis. To reduce some of this distortion, Frazier *et al.* [39] suggested a filter that adapts itself both globally and locally to the time varying nature of speech. A typical impulse response of Frazier's adaptive filter is shown in Fig. 8(b). The spacing "T_i" in Fig. 8(b) is adapted to the local variation of the pitch periods of voiced speech. A typical algorithm for speech enhancement by adaptive comb filtering (or adaptive filtering) is shown in Fig. 9.

B. Speech Enchancement Based on Harmonic Selection

The adaptive filtering technique discussed in Section IV-A requires accurate pitch information and therefore a separate

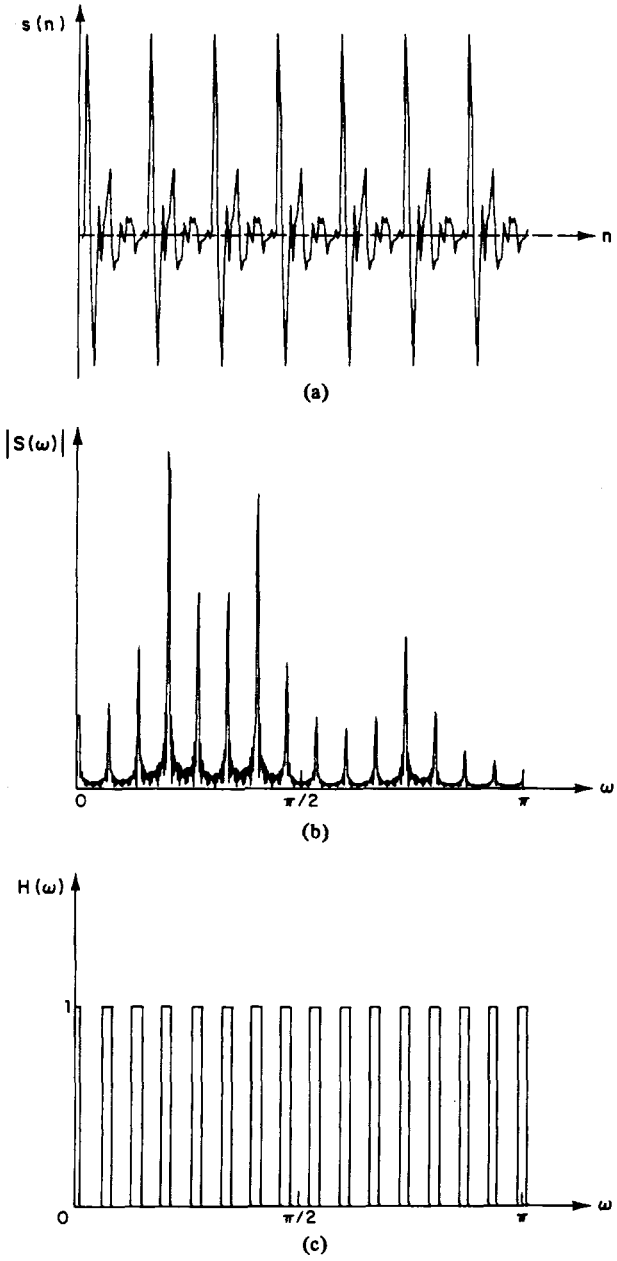

Fig. 7. (a) A periodic time waveform. (b) Magnitude spectrum of the time waveform in (a). (c) Frequency response of an ideal comb filter.

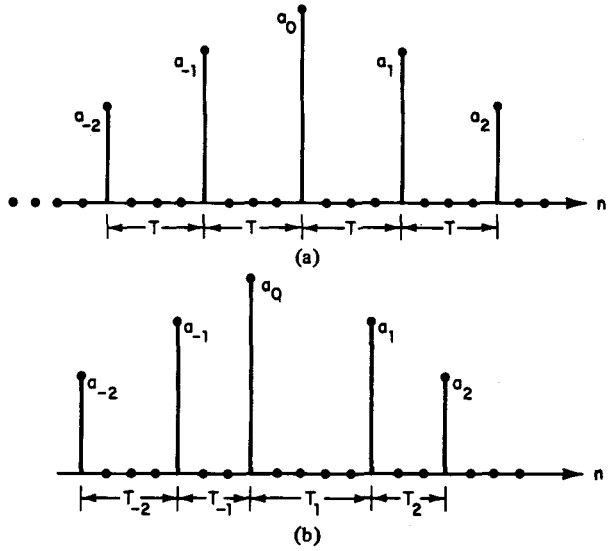

Fig. 8. (a) Impulse response of a typical adaptive comb filter by Shields [38]. (b) Impulse response of a typical adaptive filter by Frazier *et al.* [39].

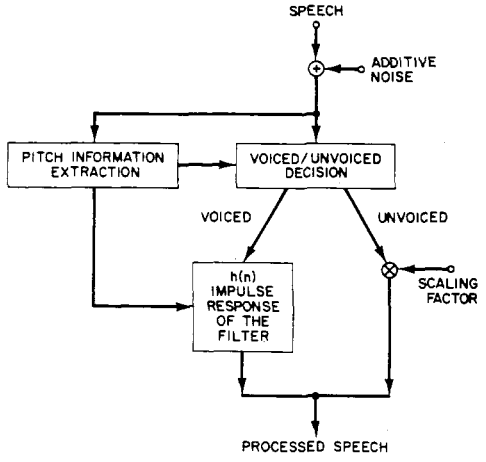

Fig. 9. A typical algorithm for speech enhancement by adaptive comb filtering or adaptive filtering.

system that estimates the pitch information is necessary. In the context of an application in which the interfering background noise is a competing speaker, Parsons [23] developed a system which is closely related to comb filtering with the pitch information obtained as an integral part of the system. Voiced speech is windowed and a high-resolution short-time spectrum is obtained. In the short-time spectrum, the periodicity of speech exhibits itself as local spectral peaks some of which are due to the main speaker and some others of which are due to a competing speaker. Parsons developed a technique in which each of the local spectral peaks in the high-resolution short-time spectrum is distinguished between the main speaker and a competing speaker. Then speech is generated based on the spectral content that corresponds to the peaks of the main speaker. Since the essence of Parsons' system is location and selection of speech harmonics of a speaker from the high-resolution spectrum of degraded speech, it can be approximately viewed as a frequency-domain implementation of a pitch information extracter and an adaptive filter.

C. Speech Enhancement Based on Adaptive Noise Canceling Techniques

A class of techniques referred to as adaptive noise canceling have been developed which are based on the availability of both the degraded signal $y(n) = s(n) + d(n)$ and a reference signal $r(n)$ which is uncorrelated with $s(n)$ but correlated with $d(n)$. A block diagram representation of such a system is shown in Fig. 10. By adaptively filtering $r(n)$ an estimate of the component $d(n)$ that is correlated with $r(n)$ is formed and subtracted from $y(n)$. Adaptive noise canceling is applicable to processing of inputs whose properties are unknown, and good performance can be achieved if a suitable reference input is available. A detailed discussion of the principles, implementations, etc. of adaptive noise cancelling can be found in [40].

As mentioned in the introduction, the discussion in this paper is restricted to systems for which the only signal available is the degraded speech and thus adaptive noise canceling as outlined above would not be applicable. However, Sambur [24] developed a system which utilizes the principles of adap-

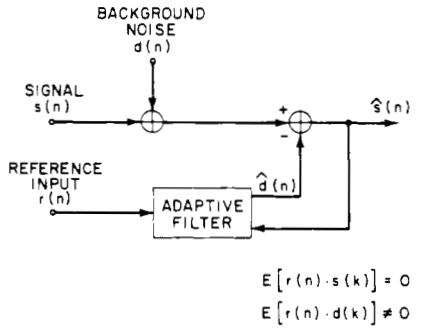

Fig. 10. An adaptive noise cancelling system.

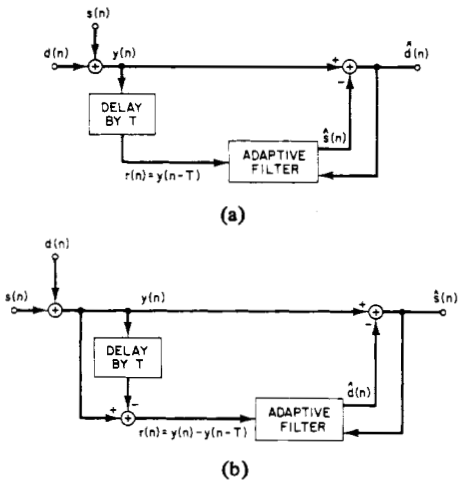

Fig. 11. (a) An adaptive noise cancelling technique for speech enhancement by Sambur [24]. (b) Another adaptive noise cancelling technique for speech enhancement.

tive noise canceling by generating a reference input, capitalizing on the periodicity of voiced speech. Specifically, let the reference input $r(n)$ be given by $r(n) = y(n - T)$, where T represents the pitch period. To the extent that periodicity is strictly observed,

$$r(n) = s(n - T) + d(n - T) = s(n) + d(n - T). \tag{29}$$

Reversing the roles of $s(n)$ and $d(n)$ in Fig. 10, $r(n)$ can be viewed as uncorrelated with $d(n)$ to the extent that the correlation of $d(n)$ is short and the adaptive filter has a short impulse response relative to the pitch period T. Since the component $s(n)$ in $r(n)$ is identical to the $s(n)$ in the primary input $y(n)$, the output of the adaptive filter in Fig. 10 would correspond to an estimate of $s(n)$. The adaptive noise cancelling technique proposed by Sambur is shown in Fig. 11(a). An alternative approach to Sambur's technique is shown in Fig. 11(b). In the figure, a reference input $r(n)$ is specified as

$$r(n) = y(n) - y(n - T). \tag{30}$$

To the extent that periodicity is strictly observed,

$$r(n) = s(n) + d(n) - s(n - T) - d(n - T) = d(n) - d(n - T). \tag{31}$$

Then $r(n)$ is uncorrelated with $s(n)$ but is highly correlated with $d(n)$ thus satisfying the condition for adaptive noise canceling.

The adaptive noise canceling technique in Fig. 11(b) can be related to comb filtering discussed in Section IV-A. Specifically, if we assume that the adaptive process has converged and the adaptive filter is short enough so that $d(n)$ can be assumed to be uncorrelated with $d(n - T)$, the results in [40] can be used to show that the frequency response of the filter is given by

$$H(\omega) = \frac{P_d(\omega)}{P_d(\omega) + P_{d'}(\omega)} \tag{32}$$

where $d'(n) = d(n - T)$. Since $P_d(\omega) = P_{d'}(\omega)$, $H(\omega)$ in (32) equals 1/2 and $\hat{d}(n)$ in the figure is given by

$$\hat{d}(n) = \tfrac{1}{2}\, y(n) - \tfrac{1}{2}\, y(n - T). \tag{33}$$

From (33) and Fig. 11(b),

$$\hat{s}(n) = \tfrac{1}{2}\, y(n) + \tfrac{1}{2}\, y(n - T). \tag{34}$$

Equation (34) is the result obtained by a comb filter whose impulse response equals $\frac{1}{2}\,\delta(n) - \frac{1}{2}\,\delta(n - T)$.

In this section, we have discussed various speech enhancement techniques which capitalize on the periodicity of voiced speech. Depending on how the periodicity of voiced speech is specifically exploited, different techniques have been developed. All of them, however, have the common feature that they are based only on the periodicity of voiced speech and require accurate pitch information. Techniques for extracting the pitch information from noisy speech will be discussed in Section VI. Some performance evaluation results and potential advantages and disadvantages of the techniques discussed in this section will be presented in Section VII.

V. Speech-Enhancement Techniques Based on a Speech Model

A digital model of sampled speech that has been used in a number of practical applications and has a basis [6]-[8] in the physics of speech production system was shown in Fig. 1. In the model, the excitation source is either a quasi-periodic train of pulses for voiced sounds or random noise for unvoiced sounds. The digital filter represents the effects of the vocal tract, lip radiation, and, for voiced sounds, the glottal source. Since the vocal tract changes in shape as a function of time, the digital filter in Fig. 1 is in general time varying. However, over a short interval of time, the digital filter may be approximated as a linear time invariant system. Many systems which capitalize on the underlying speech model discussed above have been proposed in the literature for speech enhancement and in this section we discuss some of those techniques.

In the speech enhancement technique based on an underlying speech model, the parameters of the speech model are first estimated and then speech is generated based on the estimated parameters. The parameters of the model consist of the source parameters (pitch information) and the system parameters (vocal-tract information). The problem of estimating the source parameters from noisy speech will be discussed in Section VI, where we discuss techniques for bandwidth compression of noisy speech, and in this section we consider techniques for estimating the system parameters. Given the estimated parameters of a speech model, speech can be generated by a synthesis system based on the same underlying speech model or by designing a filter with the estimated speech model parameters and then filtering the noisy speech. The former approach requires both the source and system parameters while the latter approach generally requires only the system parameters as will be discussed later.

The techniques to estimate the system parameters of a speech model, of course, depend on the specific model assumed. Even for the same speech model, however, there are again a variety of different techniques that may be used in estimating the model parameters. In Section V-A, we discuss speech enhancement techniques based on an all-pole model of the vocal tract and in Section V-B, techniques based on a pole-zero model of the vocal tract. In Section V-C, we discuss techniques based on nonparametric speech models.

A. Speech Enhancement Techniques based on an All-Pole Model of Speech

In an all-pole model of speech, the transfer function $V(z)$ in Fig. 1 is modeled on a short-time basis as all-pole of the form

$$V(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \tag{35}$$

where "p" represents the order of all-pole model. Thus on a short-time basis the speech waveform $s(n)$ is assumed to satisfy a difference equation of the form

$$s(n) = \sum_{k=1}^{p} a_k \cdot s(n-k) + u(n) \tag{36}$$

where $u(n)$ is a pulse train for voiced speech or random noise for unvoiced speech.

Equation (36) is sometimes referred to as an autoregressive model or as a linear prediction model since the current sample $s(n)$ can be viewed as being predicted from a linear combination of past samples of $s(n)$ with an error of $u(n)$. For notational convenience, the all pole parameters will be denoted in vector form as

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_p \end{pmatrix} \tag{37}$$

The problem of estimating a given a segment of $s(n)$ has been considered extensively [41], [42] in the literature. In the absence of background noise, many different approaches [41] to estimate a lead to solving essentially the same set of linear equations of the form

$$R \cdot a = r \tag{38}$$

where R is a $p \times p$ matrix and r is a $p \times 1$ matrix. Depending on how the matrices R and r are specifically obtained from $s(n)$, equation (38) is referred to as either the correlation or covariance method of linear prediction analysis. The principal advantages of the correlation method are that R in (38) is a Toeplitz matrix so that particularly efficient algorithms (43) to solve (38) exist and the resulting all-pole coefficients are guaranteed to be stable.

The problem of estimating the all-pole parameters from the noisy speech is a much more difficult problem and different approaches generally lead to different results. One approach is to simply solve (38) for all-pole parameters a where the components of R and r are estimated accounting for the presence of noise. In the correlation method of linear prediction analysis, the components of R and r consist of the first $p+1$ points of the correlation of $s_w(n)$ representing $s(n)$ multiplied by a time-limited window $w(n)$ as introduced in Section III-A. The Fourier transform of the correlation of $s_w(n)$ is $|S_w(\omega)|^2$ and in Section III we have discussed various techniques to estimate $|S_w(\omega)|$ from the noisy speech. Then one approach would be to estimate $|S_w(\omega)|^2$ from the noisy speech by one of the techniques discussed in Section III, form R and r from the inverse transform of this estimate and then solve for a in (38). Techniques to estimate the all-pole coefficients in this way have been considered by Magill and Un [28], Kobatake *et al.* [44], and Lim [18].

A more theoretical approach to the problem of estimating the all pole coefficients a is to use well-known parameter estimation rules. Before we discuss this approach in greater detail, we review very briefly the general principles of parameter estimation.

Let A and R denote the parameter space and the observation space, and assume that there is a probabilistic mapping between these spaces with a point α in the parameter space mapped to a point r in the observation space. The parameter estimation problem is to estimate the value of α from the observation r, using some estimation rule. The three estimation rules known as maximum likelihood (ML), maximum *a posteriori* (MAP), and minimum mean-square error (MMSE) estimation have many desirable properties and thus have been studied [45], [46] extensively in the literature. For nonrandom parameters, the ML estimation rule is often used. In the ML estimation, the parameter value is chosen such that the chosen value most likely resulted in the observation r. Thus the value of α is chosen such that $p_{R|A}(r|\alpha)$, the probability density function of R conditioned on A, is maximized at the observed r and the chosen value of α. The MAP and MMSE estimation rules are commonly used for parameters that can be considered as random variables whose *a priori* density function is known. In the MAP estimation rule, the parameter value is chosen such that the *a posteriori* density $p_{A|R}(\alpha|r)$ is maximized at the observed r and the chosen value of α. ML and MAP estimation rules lead to identical estimates of the parameter value when the *a priori* density of the parameter in the MAP estimation rule is assumed to be flat over the parameter space. For this reason, the ML estimation rule is often viewed as a special case of the MAP estimation rule. In the MMSE estimation rule $\hat{\alpha}(R)$, the estimate of α, is obtained by minimizing the mean-square error $E[(\hat{\alpha}(R) - \alpha)^2]$. The MMSE estimate of α is given by $E[\alpha|r]$, the *a posteriori* mean of α given r. Therefore, when the maximum of the *a posteriori* density function $p_{A|R}(\alpha|r)$ coincides with its mean, the MAP estimation and MMSE estimation rules lead to identical estimates.

Lim and Oppenheim [29] have considered estimation of the all-pole coefficients a using MAP estimation, thus maximizing $p(a|y)$ where[3] y represents the samples of noisy speech with the assumption that the excitation is white Gaussian noise. The approach was motivated partly by the fact [29], [47] that in the absence of background noise the MAP estimation procedure with white Gaussian noise excitation leads to the correlation method of linear prediction analysis which has

[3] For a more accurate representation, a probability density function $p_x(\cdot)$ and the density function evaluated at $x = x_0$ should be distinguished. For notational convenience, $p(x_0)$ will be used in both cases and the distinction will be left to the context in which it is used.

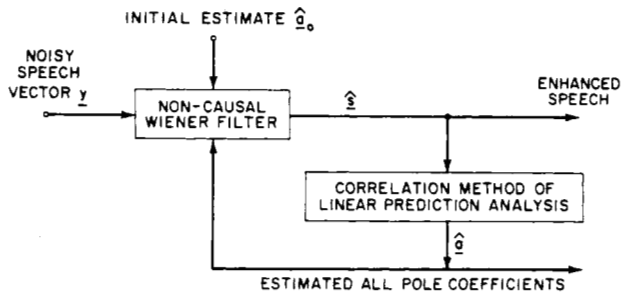

Fig. 12. Linearized MAP (LMAP) algorithm for estimation of the all pole parameters a and the speech vector s from the noisy speech vector y.

been successful in the analysis of both voiced and unvoiced speech. In the presence of background noise, the MAP parameter estimation rule leads to solving a set of nonlinear equations (29). However, if a is estimated by maximizing $p(a, s|y)$, where s represents the samples of noise-free speech, then an iterative algorithm which requires solving only sets of linear equations can be developed. The iterative algorithm, referred to as linearized MAP (LMAP) begins with an initial estimate $\hat{a}_0$ and then estimates s as $E[s|\hat{a}_0, y]$. With this estimate of s, a new estimate $\hat{a}_1$ is obtained by the correlation method of linear prediction analysis. With the new $\hat{a}_1$, the above procedure is repeated to obtain a newer estimate $\hat{a}_2$. It can be shown (29) that estimating s as $E[s|\hat{a}_i, y]$ is a linear problem and further that the above iterative procedure increases $p(a, s|y)$ in each iteration.

If an infinite amount of data is assumed to be available, it can be shown that estimating s as $E[s|\hat{a}_i, y]$ is equivalent to filtering the noisy speech with a noncausal Wiener filter whose frequency response is given by

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \tag{39}$$

with

$$P_s(\omega) = \frac{g^2}{\left|1 - \sum_{k=1}^{p} a_k \cdot \exp(-j\omega k)\right|^2} \tag{40}$$

where a_k in (40) corresponds to $\hat{a}_i$ and g^2 represents the gain in the excitation. A typical LMAP algorithm with the assumption of an infinite amount of data is shown in Fig. 12. As is clear from the figure, the approach based on maximizing $p(a, s|y)$ estimates not only the all-pole coefficients but the noise-free speech vector s. Thus either $\hat{s}$ can be utilized as the estimate of $s(n)$, or the coefficients $\hat{a}$ can be used to synthesize an estimate of $s(n)$.

In the LMAP algorithm, when a is estimated from $\hat{s}$ by the correlation method of linear prediction analysis, the values $\hat{s}$ are used to form the short-time correlation which consists of components of the form of $s(i) \cdot s(j)$. The LMAP algorithm estimates $s(i)s(j)$ by

$$\hat{s(i) \cdot s(j)} = E[s(i)|\hat{a}, y,] \cdot E[s(j)|\hat{a}, y]. \tag{41}$$

As an alternative, $s(i) \cdot s(j)$ may be estimated directly by

$$\hat{s(i) \cdot s(j)} = E[s(i) \cdot s(j)|\hat{a}, y]. \tag{42}$$

An iterative algorithm based on (42) has been referred to [29] as revised LMAP (RLMAP) algorithm. It can be shown that estimating $s(i) \cdot s(j)$ using (42) again requires solving only a set of linear equations and as with the LMAP algorithm the assumption of infinite data leads to a computationally simple procedure which has a frequency-domain representation. Furthermore, it can be shown [18], [29] that each iteration in the RLMAP algorithm increases $p(a|y)$ instead of $p(a, s|y)$, thus corresponding to a true MAP parameter estimation rule.

As an alternative approach to estimate the all-pole parameters from a noisy observation, we may model the noisy all-pole process by a pole-zero process, estimate the poles and zeros, and then identify the all-pole parameters from the estimated poles and zeros. Specifically, if we assume the excitation in an all-pole process is white Gaussian noise and the additive noise is also white Gaussian uncorrelated with the excitation, then it can be shown [48] that the noisy all-pole process can be modeled by a pole-zero system whose poles are identical to those of the original all-pole system. By using a pole-zero parameter estimation technique, the poles and zeros of the pole-zero system may be estimated and the resulting poles may be identified as the poles of the original all-pole process. This approach has been applied by Done and Rushforth [30] to estimate all-pole parameters from a noisy time series, and may be applied to estimate the all-pole parameters from noisy speech.

In the above we have discussed several approaches to estimating the parameters in all-pole model of the vocal tract. In the LMAP algorithm, the noise-free speech is estimated in the process of estimating the all-pole parameters and thus the estimate of noise-free speech can be directly used as the output of the enhancement system. In other approaches, however, speech has to be generated from the estimated all-pole parameters. One way to generate speech is to use a speech synthesis system based on the same underlying speech model used in the analysis. This approach requires an estimation of the source parameters. An alternative approach which does not require an estimation of the source parameters is to form $P_s(\omega)$ in (40) from the speech model parameters and then form an optimum filter $H(\omega)$ as in (21). Then speech can be generated by filtering the noisy speech. If the filtering is performed in the same manner as in Section III-B, i.e., $H(\omega)$ applied to $Y_w(\omega)$ to obtain the estimate $\hat{S}_w(\omega)$, the techniques discussed in this section again can be viewed as a particular method of estimating the short-time spectral amplitude of speech discussed in Section III. The difference lies in the fact that the techniques discussed in this section were developed by attempting to capitalize on a particular speech model.

B. Speech-Enhancement Techniques Based on a Pole-Zero Model of Speech

Even though the all-pole model of speech has been used in many speech communication problems, it is known [7], [8] that a variety of sounds can be more adequately modeled by a pole-zero system. In a pole-zero model of speech, the transfer function $V(z)$ in Fig. 1 is modeled on a short-time basis to be of the form

$$V(z) = \frac{\sum_{k=0}^{q} b_k \cdot z^{-k}}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \tag{43}$$

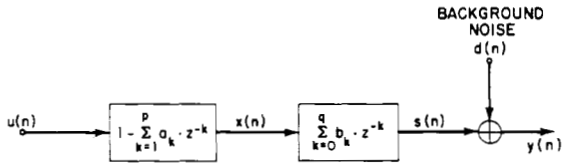

Fig. 13. A noisy speech model based on a pole-zero speech model.

where "q" represents the order of zeros. Thus on a short-time basis the speech waveform $s(n)$ is assumed to satisfy an autoregressive moving average difference equation of the form

$$s(n) = \sum_{k=1}^{p} a_k \cdot s(n-k) + \sum_{k=0}^{q} b_k \cdot u(n-k) \qquad (44)$$

where $u(n)$ represents the source excitation, and $\boldsymbol{a}$ and $\boldsymbol{b}$ are the system parameters of the model. An alternative representation of (44) is

$$x(n) = \sum_{k=1}^{p} a_k \cdot x(n-k) + u(n) \qquad (45a)$$

and

$$s(n) = \sum_{k=0}^{q} b_k \cdot x(n-k). \qquad (45b)$$

This corresponds to the overall system being represented as the cascade of an all-pole and an all-zero model as indicated in Fig. 13.

In general, estimating the zero parameters $\boldsymbol{b}$ in the presence of noise is a very difficult problem since zeros are much more easily masked by the background noise than poles. Nevertheless, techniques similar to those discussed in Section V-A have been developed to estimate the zeros in the presence of noise. One approach is to enhance speech first by the techniques discussed in Section III and then use available pole-zero parameter estimation techniques [49]-[52] applicable to noise-free signals.

Another approach to the problem of estimating the model parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ is to use well known parameter estimation rules. Musicus and Lim [31] and Musicus [32] considered using the MAP estimation rule and have shown that the iterative algorithms discussed in Section V-A for an all-pole model can be generalized to a pole-zero model. Specifically, the LMAP algorithm can be generalized by attempting to maximize $p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{x}, |y)$ where $\boldsymbol{x}$ represents the samples of $x(n)$ in (45) and Fig. 13. The generalized algorithm begins with an initial estimate $\hat{\boldsymbol{a}}_0$ and $\hat{\boldsymbol{b}}_0$, from which the estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ is formed as $\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\hat{\boldsymbol{a}}_0, \hat{\boldsymbol{b}}_0, y]$. With this estimate of $\boldsymbol{x}$, a new estimate $\hat{\boldsymbol{a}}_1$ and $\hat{\boldsymbol{b}}_1$, is obtained as $\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{b}}_1 = E[\boldsymbol{a}, \boldsymbol{b}, |\hat{\boldsymbol{x}}]$. With the new $\hat{\boldsymbol{a}}$, and $\hat{\boldsymbol{b}}$, the above procedure is repeated to obtain an updated estimate $\hat{\boldsymbol{a}}_2$ and $\hat{\boldsymbol{b}}_2$. It can be shown (32) that the steps discussed above involve solving only sets of linear equations and further that the above iterative procedure increases $p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{x}|y)$ in each iteration.

In the generalized LMAP algorithm discussed above, when $\boldsymbol{a}$ and $\boldsymbol{b}$ are estimated from $\hat{\boldsymbol{x}}$, the values $\hat{\boldsymbol{x}}$ are used to form products of the form $x(i) \cdot x(j)$. The generalized LMAP algorithm estimates $x(i) \cdot x(j)$ as

$$x(i) \hat{\cdot} x(j) = E[x(i)|\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, y] \cdot E[x(j)|\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, y]. \qquad (46)$$

As an alternative, $x(i) \cdot x(j)$ may be estimated directly as

$$x(i) \hat{\cdot} x(j) = E[x(i) \cdot x(j)|\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, y]. \qquad (47)$$

As with the all-pole case, an iterative algorithm based on (47) increases $p(\boldsymbol{a}, \boldsymbol{b}|y)$ in each iteration (32). In both the generalized LMAP and RLMAP algorithms, an infinite data assumption leads to a computationally simple procedure which has a frequency domain representation. Generation of speech from the estimated model parameters is essentially the same as in the all-pole model case discussed in Section V-A.

C. Speech-Enhancement Techniques Based on a Nonparametric Model of Speech

In Sections V-A and V-B, we have considered speech enhancement techniques based on a parametric model of the vocal-tract transfer function $V(z)$. Nonparametric representations for $V(z)$ such as homomorphic analysis of speech can also be considered (53). For a nonparametric representation of $V(z)$, it is the impulse response $v(n)$ which is estimated rather than the model parameters. Two specific speech enhancement techniques which are based on a nonparametric model of speech are a system developed by Miller [25] to remove record noise and the orchestral accompaniment from early recordings of Enrico Caruso, and a system by Suzuki [26], [27]. The two systems were briefly discussed in Section II.

A simple alternative approach to capitalize on a nonparametric representation of speech is to first enhance speech by any of the techniques discussed in Section III, and then estimate the impulse response by deconvolution techniques [1], [54] based on a nonparametric representation of speech such as homomorphic speech analysis [53]. A more theoretical approach to estimating the impulse response based on classical estimation rules is a much more difficult problem. Even though iterative algorithms similar to those discussed in Sections V-A and V-B can in principle be developed, relating the algorithms to an estimation rule such as MAP estimation is not an easy task.

VI. Techniques for Bandwidth Compression of Noisy Speech

Much of the discussion in the previous sections focused on the problem of processing degraded speech in preparation for listening, with the objective of improving quality, intelligibility or some other attribute. A related but distinct problem is that of processing degraded speech in preparation for coding by a bandwidth compression system. It is commonly understood that robustness is a problem in bandwidth compression of speech, specifically that performance degrades quickly [55]-[57] as the signal-to-noise ratio decreases. Thus it is important to develop techniques for bandwidth compression which specifically account for the presence of noise.

There are two basic approaches typically considered. The first, depicted in Fig. 14 corresponds to using a conventional bandwidth compression system preceded by a preprocessor to first reduce the background noise. In this case any of the variety of noise reduction systems which were discussed previously could potentially be used. A number of systems for bandwidth compression of noisy speech in the form of Fig. 14 have been implemented and evaluated. Typically, whereas the intelligibility of the output of the noise reduction system is less than that of the input, the intelligibility of the output of the

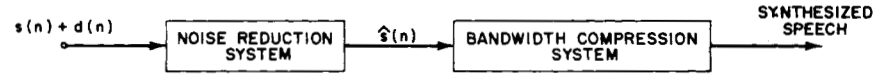

Fig. 14. Bandwidth compression of noisy speech using a noise reduction system as a preprocessor.

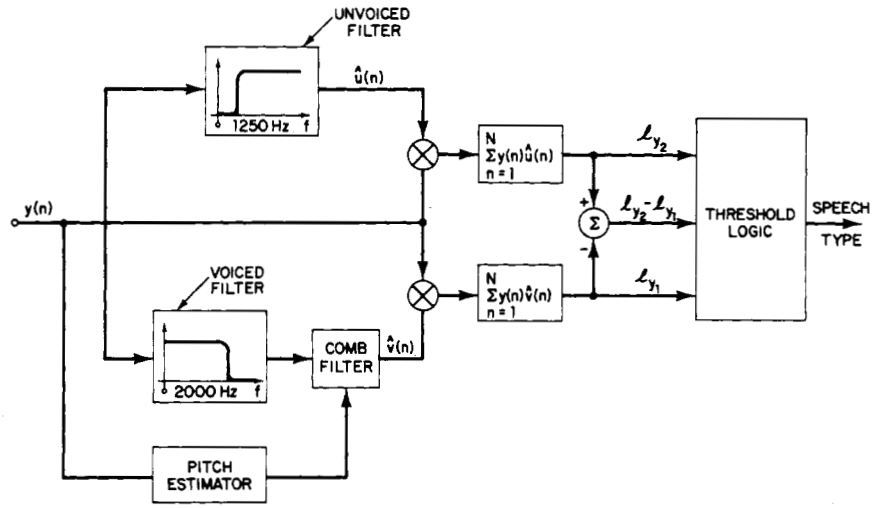

Fig. 15. A system for classification of noisy speech by McAulay [63], [64].

bandwidth compression system is higher than would be achieved if the noise recuction system were not present.

An alternative approach is to directly incorporate into the bandwidth compression system the knowledge that the model for the input signal is speech plus additive noise. For example, various systems for compression of undegraded speech are based on parametric modeling of the speech waveform [41], [42]. The parameters are coded and transmitted and at the receiver are then used to resynthesize the speech. One particularly successful form for such a system referred to as linear predictive coding (LPC) represents the speech signal in the form of Fig. 1 with the vocal-tract transfer function modeled on a short-time basis as an all-pole filter. As was discussed in Section V-A, there are available a variety of successful approaches to estimating the parameters of the vocal-tract transfer function. The remaining parameters are those used to represent the excitation function and correspond to a decision as to whether, for each segment, the speech is voiced or unvoiced, and if voiced a determination of the fundamental frequency. Again, for the case of undegraded speech, there are a variety of successful systems for estimating the excitation parameters [58]-[61].

For speech degraded by additive background noise we can, in a similar fashion, attempt to estimate the parameters. In particular, in Section V-A we discussed for degraded speech the estimation of the parameters in an all-pole model and in Section V-B the estimation of the parameters in a pole-zero model using MAP parameter estimation techniques. In the context of that discussion the parametric modeling was directed at an enhancement system. Clearly, however, the parameters can be coded with the speech resynthesized at the receiver, just as is done with conventional LPC. In addition to the resulting bandwidth compression, the system also performs as a speech enhancement system.

Another example of a speech compression system which has been modified to account for the presence of additive noise is the spectral envelope estimation vocoder developed by Paul [62]. In his speech compression system, the vocal tract transfer function is estimated by first carrying out a high-resolution spectral analysis for each speech frame. The peaks corresponding to the spectral envelope at the frequencies of the harmonics of the fundamental frequency are then located. Next, the spectrum is interpolated between these frequencies to obtain an estimate of the spectral envelope, corresponding to the vocal-tract transfer function. In the modification of the system when background noise is present, the assumed spectrum for the background noise is subtracted from the spectral envelope obtained for the degraded speech. This new estimate for the vocal tract transfer function is then used to provide the parameters for the synthesizer.

The above approaches provide several alternatives for obtaining parameters representing the vocal-tract transfer function. In general it appears to be considerably more difficult to extract excitation parameters from degraded speech. Essentially all algorithms for determination of excitation parameters with undegraded speech become seriously degraded with even moderate signal-to-noise ratios and to a large extent the estimation ot excitation parameters from noisy speech remains a current area of research. Particularly difficult and unresolved is the determination of whether a given segment of speech is voiced, unvoiced or silence. McAulay [63], [64] has proposed one system for optimum speech classification based on the principles of decision theory. The resulting system is shown in Fig. 15. This system requires an estimate of the fundamental frequency under the hypothesis that the speech is voiced. For voiced speech, one approach for determination of the fundamental frequency that has been particularly successful is the maximum likelihood pitch estimator as proposed by Wise *et al.* [65]. They formulated the problem as that of estimating an unknown periodic signal in white Gaussian noise of unknown intensity. The resulting procedure for obtaining the optimum estimate of the pitch period corresponds to constructing a bank of comb filters each tuned to a slightly different pitch period and choosing as the estimate the pitch cor-

responding to the comb filter for which the output energy is largest.

Another, somewhat different approach to obtaining an excitation for the synthesizer, which requires a higher data rate has been proposed by Magill and Un [28]. The overall system for noise reduction is based on the use of all-pole modelling of the vocal tract transfer function as outlined in Section V, with an excitation function obtained by passing the result of low-pass filtering the residual signal for the noisy speech through a nonlinear distortion to broaden its bandwidth.

VII. Performance Evaluation

The performance evaluation of the various systems discussed in this paper is a very difficult task, partly because the performance of a system may vary depending on the particular application under consideration. Some systems which improve speech quality may decrease speech intelligibility. Some systems which improve speech intelligibility in the context of bandwidth compression may decrease speech intelligibility in the context of speech enhancement. Some systems which improve speech quality when the speech degradation is due to additive random noise may not even be applicable if the degradation is due to a competing speaker.

A further complicating factor in evaluating the system performance is that the objective of various systems discussed in this paper is generally an improvement in some aspects of human perception such as an improvement in speech intelligibility or quality, or reduction of listener fatigue. Since the human perceptual domain is not well understood, a careful system evaluation requires a subjective test such as a speech intelligibility or quality test. A careful subjective test can be tedious and time consuming, and generally requires processing a large amount of data.

Because of the difficulty involved in the evaluation, only a few systems have been carefully evaluated by a subjective test for some particular environments. A few others have only been evaluated based on an objective measure such as S/N ratio improvement even though such an objective measure does not correlate well with a subjective measure. In this section, we summarize the performance evaluation that has been reported for some of the systems presented in this paper. Since the evaluation has been based on different procedures, test material, environments, etc., no attempt is made to compare individual systems. In Section VII-A, the evaluation of high-pass filtering and clipping for speech enhancement is summarized. It has been reported that this system noticeably improves intelligibility despite the fact that speech quality is seriously degraded. In Section VII-B, the evaluation of high-pass filtering for the specific phoneme /s/ and creating short pauses before plosive sounds for speech enhancement has been summarized. It is reported that this system noticeably improves speech intelligibility if tne locations of the phoneme /s/ and the plosive sounds are accurately known. In Section VII-C, the evaluation of one of the spectral subtraction techniques is summarized. In the context of speech enhancement the system does not improve speech intelligibility but improves speech quality. In the context of bandwidth compression, the system appears to improve intelligibility. In Section VII-D, the evaluation of adaptive comb filtering for speech enhancement is summarized. Here again despite an improvement in S/N ratio, the system reduces intelligibility. In Section VII-E, the evaluation of splicing of autocorrelation function (SPAC) indicating an improvement in speech quality is summarized. In Section VII-F, the evaluation of the LMAP and RLMAP techniques is summarized. The LMAP technique appears to improve speech quality both in the context of speech enhancement and bandwidth compression. Based on an objective measure, the LMAP and RLMAP techniques estimate speech synthesis parameters more accurately in the context of bandwidth compression.

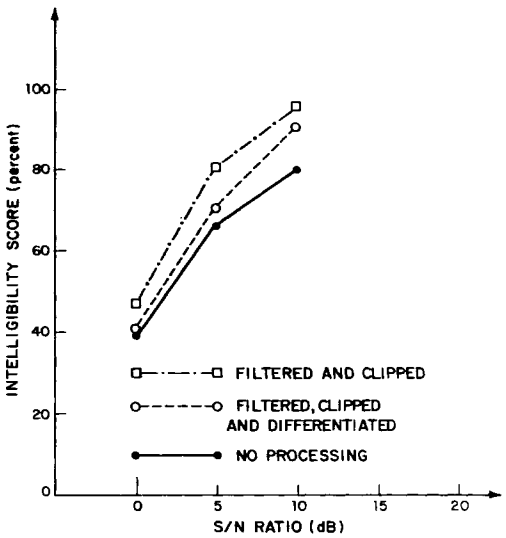

Fig. 16. Intelligibility scores of high-pass filtering and clipping, and high-pass filtering, clipping and differentiation for enhancement of speech degraded by wide-band random noise. after Thomas and Ravindran [13].

A. High-Pass Filtering and Clipping

As was discussed in Section II, high-pass filtering and clipping have been considered for speech enhancement by Thomas and Ravindran [13]. Their evaluation was based on a speech intelligibility test with the test material of Harvard PB-50 (phonetically balanced) word lists when the degradation is wide-band random noise. They also evaluated high-pass filtering, clipping and differentiation for speech enhancement. The results of their evaluation are shown in Fig. 16.

Before we discuss the results of the evaluation, we review very briefly speech intelligibility tests. In a typical speech intelligibility test [66], [67], listeners are presented with test material and asked to identify the test material or answer questions based on the test material. For example, listeners may be presented with sentences, words or syllables and asked to write the test material that they heard or choose one out of several options which most closely resembles what they heard. Alternatively, listeners may be presented with a paragraph and asked to answer questions based on the contents of the presented paragraph. From the responses of the listeners the intelligibility score, the percentage of "correct" answers based on some predetermined criterion, is computed. For a given type of degradation, the intelligibility score is generally obtained for several different levels (amounts) of degradation. The amount of degradation is represented in terms of S/N ratio. For the same type and level of degradatoin, the intelligibility score can vary considerably depending on the test procedure, test material, training of subjects, etc. Furthermore, the definition of S/N ratio employed varies from one evaluation to another. Therefore, two systems evaluated differently and possibly with a different definition of S/N ratio cannot be compared based on the intelligibility scores alone. However, it is generally established that if one system is superior to

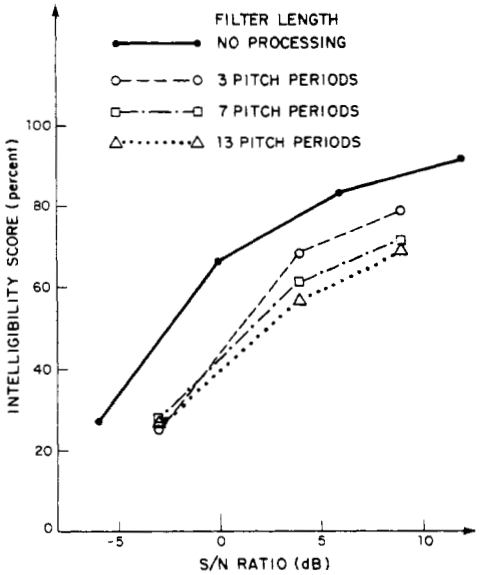

Fig. 19. Intelligibility scores of Frazier's filtering technique for enhancement of speech degraded by a competing speaker. (After Perlmutter *et al.* [73].)

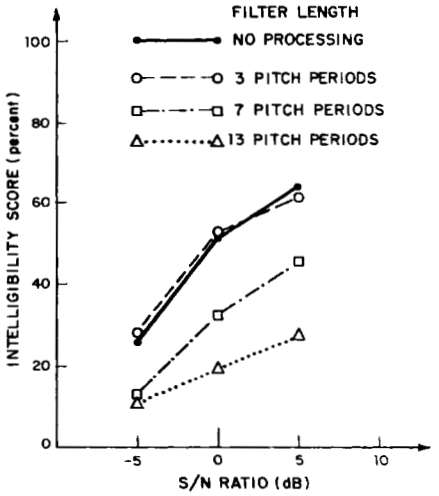

Fig. 20. **Intelligibility scores of Frazier's adaptive filtering technique improved by Lim *et al.* [74] for enhancement of speech degraded by wide-band random noise. (After Lim *et al.* [74].)**

the adaptive filtering was obtained from noise-free speech. The results of the test are shown in Fig. 19. Their results indicate that even with accurate pitch information, the adaptive filtering technique decreases intelligibility at the S/N ratios at which the intelligibility of unprocessed nonsense sentences range between 20 and 70 percent.

A modification of Frazier's adaptive filtering technique was evaluated using nonsense sentences as test material when the degradation is due to wide-band random noise [74]. The pitch information used in the processing was obtained from noise-free speech. The results of the test are shown in Fig. 20. Again, the results show that even with accurate pitch information, the adaptive filtering technique tends to decrease the intelligibility at various S/N ratios. Since in practice accurate pitch information is not available and cannot be expected to be obtained from degraded speech, the intelligibility scores will be even lower than shown in Figs. 19 and 20.

To the extent that voiced speech is periodic, the S/N ratio improvement for voiced speech using Frazier's adaptive filtering can be analytically calculated. For the modified adaptive filters [74], the S/N ratio increase is 3.5, 7, and 10 dB corresponding to the filter lengths of 3, 7, and 13 pitch periods. It is interesting to note that a higher S/N ratio increase leads to a lower intelligibility score. This is partly due to the fact that voiced speech is not strictly periodic and the periodicity assumption is more seriously violated by a filter with a longer impulse response thus causing a higher signal distortion. Despite the decrease in speech intelligibility, speech processed by an adaptive filter sounds "less noisy" due to the capability of the system to increase the S/N ratio.

E. SPAC

As was discussed in Section II, a speech enhancement system based on splicing of autocorrelation function (SPAC) was developed by Suzuki [26]. The system was evaluated by Nakatsui [75] based on a speech quality test when the degradation is due to wide-band random noise. The results of the test show that above 5 dB of S/N ratio, SPAC does not improve speech quality. In fact, at high S/N ratios, SPAC is expected to decrease speech quality since SPAC replaces one period of speech with a corresponding period of short-time autocorrelation function thus causing some speech distortion. Below about 5 dB of S/N ratio, however, some improvement in speech quality by SPAC was reported.

F. LMAP and RLMAP

The LMAP technique discussed in Section V was evaluated by Lim [18] based on a speech quality test using sentences as test material when the degradation is due to wide-band random noise. The results of the test indicate that the LMAP technique improves speech quality at various S/N ratios both in the context of speech enhancement and bandwidth compression of noisy speech.

Both the LMAP and RLMAP algorithms were evaluated by Lim [18] based on an objective measure in the context of bandwidth compression of noisy speech. In the evaluation, a number of sequences of noisy synthetic data were generated by exciting *known* all pole filters with white Gaussian noise or a train of pulses and then adding wide-band random noise at various S/N ratios. From the noisy synthetic data, all pole coefficients were estimated by the correlation method of linear prediction analysis, the LMAP and RLMAP algorithms discussed in Section V. The estimated all pole coefficients were then compared with the known all pole coefficients to form an error measure defined by

$$E = k \cdot \int_{-\pi}^{\pi} \left[\log \left| 1 - \sum_{i=1}^{p} a_i \cdot \exp(-j \cdot \omega i) \right| - \log \left| 1 - \sum_{i=1}^{p} \hat{a}_i \cdot \exp(-j \cdot \omega i) \right| \right]^2 \cdot d\omega \quad (48)$$

where "k" is a constant, a_i and $\hat{a}_i$ represent the known and estimated all pole coefficients. The error measure E defined by (48) has some correlation with perceptually important aspects of speech (41). In Fig. 21(a) is shown the error E averaged over many different sets of all pole coefficients when the excitation is white Gaussian noise. In Fig. 21(b) is shown the averaged error E when the excitation is a train of pulses. The results in Fig. 21 indicate that based on objective measure given by (48) the LMAP or RLMAP algorithm estimates the all pole coefficients more accurately than the correlation

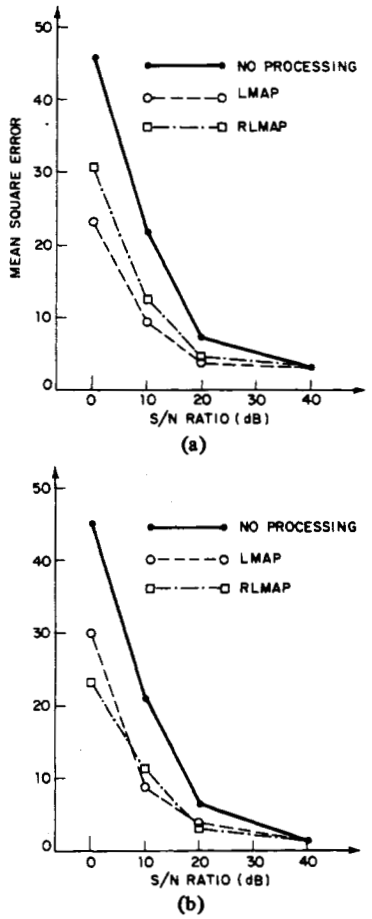

Fig. 21. Performance comparison of correlation method, LMAP and RLMAP techniques in estimating all-pole parameters from noisy Synthetic data. (After Lim [18].) (a) Random-noise excitation. (b) Pulse train excitation.

method of linear prediction at various S/N ratios when the background noise is wide-band random noise.

VIII. Conclusions

In this paper, we have attempted to survey a variety of systems for speech enhancement and to incorporate them within a common framework. As was evident in the discussion it is possible to generate an almost unlimited number of systems many of which are conceptually plausible. Furthermore many of these systems lead to an improved speech to noise ratio which is perceived as higher quality, particularly when the test material is familiar to the listener so that intelligibility is not an issue. However, almost all of these systems in fact reduce intelligibility and those that do not tend to degrade the quality. This suggests then that there remains considerable further work to be done and room for improvement.

As an additional important consideration the evaluation of an enhancement system is very much dependent on the context in which it is to be used. In some applications it is intelligibility that is of overriding importance and in others it is quality. Additionally a system may perhaps slightly reduce intelligibility but also reduce listener fatigue so that with an extended listening task intelligibility is eventually increased. To our knowledge none of the systems discussed have been evaluated in terms of their potential to reduce listener fatigue.

Essentially all of the systems considered here have their basis in a mathematically optimal procedure such as minimization of mean square error or maximization of a probability function, followed by a number of empirical variations. It is generally known that these criteria are not particularly well matched to auditory perception and it remains to develop a mathematical error criterion that strongly correlates with human perception.

An area in which speech enhancement systems have been successful is in the context of bandwidth compression. Since speech bandwidth-compression systems tend to degrade quickly in the presence of background noise, preprocessing with a speech-enhancement system prior to bandwidth compression leads to higher intelligibility after compression than would be obtained without the preprocessor. In addition as was discussed in Section VI some systems are specifically formulated as analysis-synthesis or bandwidth-compression systems with noisy inputs. Of particular difficulty in narrowband speech compression systems is the determination of excitation parameters including pitch and a voiced, unvoiced or silence decision.

We hope that the framework developed in this paper will provide the basis for further research into speech enhancement techniques and will avoid the rediscovery of existing techques. In our opinion, the problem remains an important and vital one with a need for fresh approaches and insights which we hope will emerge over the next several years.

References

[1] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1975.
[2] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, Aug. 1968.
[3] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 145–154, Feb. 1962.
[4] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun. Technol.*, vol. COM-12, pp. 162–165, Dec. 1964.
[5] J. S. Lim and A. V. Oppenheim, "Reduction of quantization noise in PCM speech coding," Tech. Note 1979-47, M.I.T.Lincoln Lab., May 31, 1979.
[6] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Lexington, MA, Mouton, 1970.
[7] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972.
[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall, 1978.
[9] I. B. Thomas, "The influence of first and second formants on the intelligibility of clipped speech," *J. Audio Eng. Soc.*, vol. 16, pp. 182–185, Apr. 1968.
[10] A. Agrawal and W. C. Lin, "Effect of voiced speech parameters on intelligibility of PB words," *J. Acoust. Soc. Amer.*, vol. 57, pp. 217–222, Jan. 1975.
[11] J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech," *J. Acoust. Soc. Amer.*, vol. 20, pp. 42–51, 1947.
[12] I. B. Thomas and R. J. Niederjohn, "Enhancement of speech intelligibility at high noise levels by filtering and clipping," *J. Audio Eng. Soc.*, vol. 16, pp. 412–415, Oct. 1968.
[13] I. B. Thomas and A. Ravindran, "Intelligibility enhancement of already noisy speech signals," *J. Audio Eng. Soc.*, vol. 22, pp. 234–236, May 1974.
[14] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 165–168, June 1968.
[15] M. R. Weiss *et al.*, "Processing speech signals to attenuate interference," presented at the IEEE Symp. Speech Recognition, Apr. 1974.
[16] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility," Nicolet Scientific Corp., Final Rep. NSC-FR/4023, Dec. 1974.
[17] R. A. Curtis and R. J. Niederjohn, "An investigation of several frequency domain processing methods for enhancing the intelligibility of speech in wideband random noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 606–609, Apr. 1978.
[18] J. S. Lim, "Enhancement and bandwidth compression of noisy speech by estimation of speech and its model parameters," Sc.D.

dissertation, Dept. Elec. Eng. and Comput. Sci., Massachusetts Inst. Technol. Cambridge, Aug. 1978.
[19] —, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-26, pp. 471–472, Oct. 1978.
[20] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-29, pp. 113–120, Apr. 1979.
[21] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 208–211, Apr. 1979.
[22] R. D. Preuss, "A frequency domain noise cancelling preprocessor for narrowband speech communications systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 212–215, Apr. 1979.
[23] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, Oct. 1976.
[24] M. R. Sambur, "Adaptive noise cancelling for speech signals," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-26, pp. 419–423, Oct. 1978.
[25] N. J. Miller, "Recovery of singing voice from noise by synthesis," Thesis Tech. Rep., ID UTEC-CSC-74-013, May 1973, Univ. Utah, Computer Science Library, Salt Lake City, UT.
[26] J. Suzuki, "Speech processing by splicing of autocorrelation function," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 113–116, Apr. 1976.
[27] —, "Speech processing system by use of short-time cross-correlation function," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 24–27, May 1977.
[28] D. T. Magill and C. K. Un, "Wide-band noise reduction of noisy speech," in *Conf. Abstract 92nd Meet. Acoust. Soc. Amer.*, *J. Acoust. Soc. Amer.*, vol. 60, Suppl. no. 1, p. S107, Fall 1976.
[29] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-26, pp. 197–210, June 1978.
[30] W. J. Done and C.-K. Rushforth, "Estimating the parameters of a noisy all-pole process using pole-zero modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, pp. 228–231, Apr. 1979.
[31] B. R. Musicus and J. S. Lim, "Maximum likelihood parameter estimation of noisy data," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, pp. 224–227, Apr. 1979.
[32] B. R. Musicus, "An iterative technique for maximum likelihood parameter estimation on noisy data," S. M. thesis, Dep. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, May 1979.
[33] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision maximum likelihood noise suppression filter," Tech. Note 1979-31, M.I.T. Lincoln Lab., Lexington, MA, June 13, 1979.
[34] M. W. Callahan, "Acoustic signal processing based on the short-time spectrum," Ph.D. dissertation, Dep. Comput. Sci., Univ. Utah, Salt Lake City, Mar. 1976.
[35] W. K. Pratt, *Digital Image Processing*. New York: Wiley, 1978.
[36] H. C. Andrews, B. R. Hunt, *Digital Image Restoration*. Englewood Cliffs, NJ: Prentice-Hall, 1977.
[37] E. R. Cole, "The removal of unknown image blurs by homomorphic filtering," Ph.D. dissertation, Dep. Elec. Eng., Univ. Utah, Salt Lake City, June 1973.
[38] U. C. Shields, Jr., "Separation of added speech signals by digital comb filtering," S.M. thesis, Dep. Elec. Eng., Massachusetts Inst. Technol., Cambridge, 1970.
[39] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. IEEE Int. Conf. Acoustic, Speech, and Signal Processing*, pp. 251–253, Apr. 1976.
[40] B. Widrow, J. R. Glover, Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. S. Zeidler, E. Dong, Jr., and R. C. Goodlin, "Adaptive noise cancelling; Principles and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.
[41] J. D. Markel, A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer, 1976.
[42] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
[43] N. Levinson, "The Wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 125, pp. 261–278, 1947.
[44] H. Kobatake, J. Inari, and S. Kakuta, "Linear predictive coding of speech signals in a high ambient noise environment," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 472–475, Apr. 1978.
[45] H. L. Van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968.
[46] R. Eykhoff, *System Identification: Parameter and State Estimation*. New York: Wiley, 1974.
[47] F. Itakura and B. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," presented at the 6th Int. Conf. Acoust., paper C-5-5, Tokyo, Japan, Aug. 1968.
[48] M. Pagano, "Estimation of models of autoregressive signal plus white noise," *Ann. Statist.*, vol. 2, pp. 99–108, 1974.
[49] J. Durbin, "Efficient estimation of parameters in moving-average models," *Biometrika*, vol. 46, pp. 306–316, 1959.
[50] J. L. Shanks, "Recursive filters for digital processing," *Geophys.*, vol. 32, pp. 33–51, 1967.
[51] A. V. Oppenheim, G. E. Kopec, and J. Tribolet, "Signal analysis by homomorphic prediction," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-24, pp. 327–332, Aut. 1976.
[52] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-25, pp. 229–234, June 1977.
[53] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 221–226, June 1968.
[54] J. S. Lim, "Spectral root homomorphic deconvolution system," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-29, pp. 223–233, June 1979.
[55] B. Gold, "Robust speech processing," Tech. Note 1976-6, M.I.T. Lincoln Laboratory, Lexington, MA, AD-A021899/0, Jan. 27, 1976.
[56] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantization noise or additive white noise," *IEEE Trans Acoust., Speech, Signal Proc.*, vol. ASSP-24, pp. 488–494, Dec. 1976.
[57] C. F. Teacher and D. Coulter, "Performance of LPC vocoders in a noisy environment," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 216–219, Apr. 1979.
[58] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1967.
[59] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, 1969.
[60] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, 1972.
[61] L. R. Rabiner, M. J. Cheng, A. E. Rosenbert, and C. A. McGonegal, "A comparative perfomance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 399–418, Oct. 1976.
[62] D. B. Paul, "A robust vocoder with pitch-adaptive spectral envelope estimation and an integrated maximum likelihood pitch estimator," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 64–67, Apr. 1979.
[63] R. J. McAulay, "Optimum speech classification and its application to adaptive noise cancellation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 425–428, Apr. 1977.
[64] R. J. McAulay, "Design of a robust maximum likelihood pitch estimator for speech in additive noise," Tech. Note 1979-28, M.I.T. Lincoln Lab., June 11, 1979.
[65] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-24, pp. 418–423, Oct. 1976.
[66] J. P. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58, pp. 955–991, 1948.
[67] C. E. Williams, M. H. L. Hecker, K. N. Stevens, and B. Woods, "Intelligibility test methods and procedures for the evaluation of speech communication systems," Final Rep., BBN, Inc., Cambridge, MA, Dec. 1966, AD-646781.
[68] M. H. Hecker and N. Guttman, "A survey of methods for measuring speech quality," *J. Audio Eng. Soc.*, vol. 15, pp. 400–413, 1967.
[69] "IEEE Recommended Practice for Speech Quality Measurements," IEEE Standards Committee Rep. 297, June 1969.
[70] L. H. Nakatani and K. D. Dukes, "A sensitive test of speech communication quality," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1083–1092, 1973.
[71] W. D. Voiers *et al.*, "Research on diagnostic evaluation of speech intelligibility," Final Rep. AFSC Contract R19628-70-C-0182, 1973.
[72] S. Meister, "The diagnostic rhyme test (DRT): An air force implementation," RADC-TR-78-129, May 1978, AD-A060917.
[73] Y. M. Purlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, "Evaluation of a speech enhancement system," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 212–215, May 1977.
[74] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust., Speech, Signal, Proc.*, vol. ASSP-26, pp. 354–358, Aug. 1978.
[75] M. Nakatsui, "Subjective evaluation of SPAC in improving quality of noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech Sig. Proc.*, Washington, D.C., pp. 467–470. April 1979.