SPEECH ENHANCEMENT

OVERVIEW

In many speech communication settings, the presence of background interference can cause the quality or intelligibility of speech to degrade. When a speaker and listener communicate in a quiet environment or across a clean data channel, information exchange is easy and accurate. However, if the environment is noisy, the listener's ability to understand what is spoken is reduced. The quality of speech can also be influenced in data conversion (microphone), transmission (noisy data channels), or reproduction (loudspeakers, headphones). The purpose of many enhancement algorithms is to reduce background noise, improve speech quality, or suppress channel or speaker interference. In this article, the general problem of speech enhancement is addressed with emphasis on algorithms relating to removal of additive background noise for improving speech quality. In our discussion, background noise will refer to any additive broadband noise component (examples include white Gaussian noise, aircraft cockpit noise, or machine noise in a factory environment). Other speech-processing areas that are sometimes included in a discussion of speech enhancement include suppression of distortion from voice-coding algorithms, suppression of a competing speaker in a multispeaker setting, enhancing speech as a result of a deficient speech production system (examples include speakers with pathology or divers breathing a helium–oxygen mixture), or enhancing speech for hearing-impaired listeners. Since the range of possible applications is broad, we will generally limit our discussion to enhancement algorithms directed at improving speech quality in additive broadband noise for speakers and listeners with normal production and auditory systems. Other sources that consider speech enhancement include the edited text by Lim (1), Chap. 8 in the text by Deller, Proakis, and Hansen (2), and survey articles by O'Shaughnessy (3) and Ephraim (4).

Speech-Enhancement Performance Criteria

As in any engineering problem, it is useful to have a clear understanding of the objectives and our ability to measure system performance in achieving those objectives. When we consider noise reduction, we normally think of improving a signal-to-noise ratio (SNR). It is important to note, however, that this may not be the most appropriate performance criterion for speech enhancement. All listeners have an intuitive understanding of speech quality, intelligibility, and listener fatigue. However, these areas are not easy to quantify in most speech-enhancement applications since they are based on *subjective* evaluation of the processed signal. A good overview of subjective quality testing methods and objective speech quality measures can be found in Quackenbush, Barnwell, and Clements (5) or Ref. 2. One of the more successful quality measures based on the speech magnitude spectrum is the Itakura-Saito (IS) distance [Chu and Messerschmitt (6); Itakura (7)]. This measure is briefly described here, since it will be used as an evaluation tool for a number of enhancement algorithms in this article. The symmetric form of this measure is based on the dissimilarity between all-pole speech spectra of the original $1/A(\boldsymbol{a}_s, \omega)$ and degraded (or enhanced) waveforms $1/A(\hat{\boldsymbol{a}}_s, \omega)$ as

$$d_j(\boldsymbol{a}_s, \hat{\boldsymbol{a}}_s) = \int_{-\pi}^{+\pi} \left[e^{v(\omega)} - v(\omega) - 1\right] \frac{d\omega}{2\pi} \tag{1}$$

$$v(\omega) = \log\left(\frac{1}{|A(\boldsymbol{a}_s, \omega)|^2}\right) - \log\left(\frac{1}{|A(\boldsymbol{a}_s, \omega)|^2}\right) \tag{2}$$

where $\boldsymbol{a}_s$ and $\hat{\boldsymbol{a}}_s$ are the all-pole model parameters from the gain-normalized original and coded speech spectra. The symmetric IS measure is obtained as follows,

$$d_{\mathrm{IS}(j)} = \frac{1}{2}\{d_j(\boldsymbol{a}_s, \hat{\boldsymbol{a}}_s) + d_j(\hat{\boldsymbol{a}}_s, \boldsymbol{a}_s)\}. \tag{3}$$

The measure has been shown to assign a large weight when error is due to differences in the shape of spectral peaks and a smaller weight for error in spectral valleys. This is desirable, since the auditory system is more sensitive to errors in formant peaks than to spectral valleys between peaks. If the numerical value of $d_{\mathrm{IS}(j)}$ is zero, then the speech signal $\hat{\boldsymbol{s}}_j$ at frame j is the same as the noise-free reference $\boldsymbol{s}_j$.

Now, consider the speech signal plotted in Fig. 1. Figure 1(a) shows the clean speech waveform "You can talk the game, but can you play the game?" Figure 1(b) shows the sentence degraded with additive white Gaussian noise with an overall SNR of 5 dB. In Fig. 1(c), the IS objective speech quality measure is shown across time. The IS measure here assesses the level of distortion for each frame location in time. Since the energy and frequency content of the speech signal varies across time, due to the sequence of phonemes needed to produce the sentence, the impact of background distortion will also vary. This variable level of speech distortion (or quality) is reflected in the changing IS measure. Essentially, the area under the IS plot versus time reflects the distortion level. We can see that while the average SNR level is 5 dB, noise will effect some sounds (e.g., *stops* /t/, /g/; or fricatives / s/, /z/) more than others (e.g., vowels /e/, /I/). Therefore, when considering the performance of speech-enhancement methods, it is necessary to remember the issue of a nonuniform impact of the noise on speech quality across time.

Classification of Speech-Enhancement Methods

Speech-enhancement systems are classified into two broad classes: those based on stochastic process models of speech and those based on perceptual aspects of speech. Systems based on stochastic process models rely on a given mathematical criterion (i.e., mean-square error, SNR).Systems based on perceptual criteria attempt to improve aspects important in human perception. For example, one technique may concentrate on improving the quality of consonants, since consonants are known to be important for intelligibility in a manner disproportionate to overall signal energy (this is particularly important for hearing-aid design). Figure 2 illustrates a flow diagram of speech-enhancement applications and potential sources of distortion. Distortion may consist of background noise, competing speakers, room reverberation, voice-coder distortion, or channel interference. Speech enhancement could then be used to improve characteristics in the speech signal prior to human listening or other speech processing algorithms (e.g., speech recognition).

In the area of digital hearing aids, a number of recent studies have shown promise using better speech versus noise detection and microphone arrays [Kates and Weiss (8)], wavelet-based spectral attenuation [Whitmal, Rutledge, and Cohen (9)] and real-time DSP (digital signal processing) development using a previously formulated spectral estimator

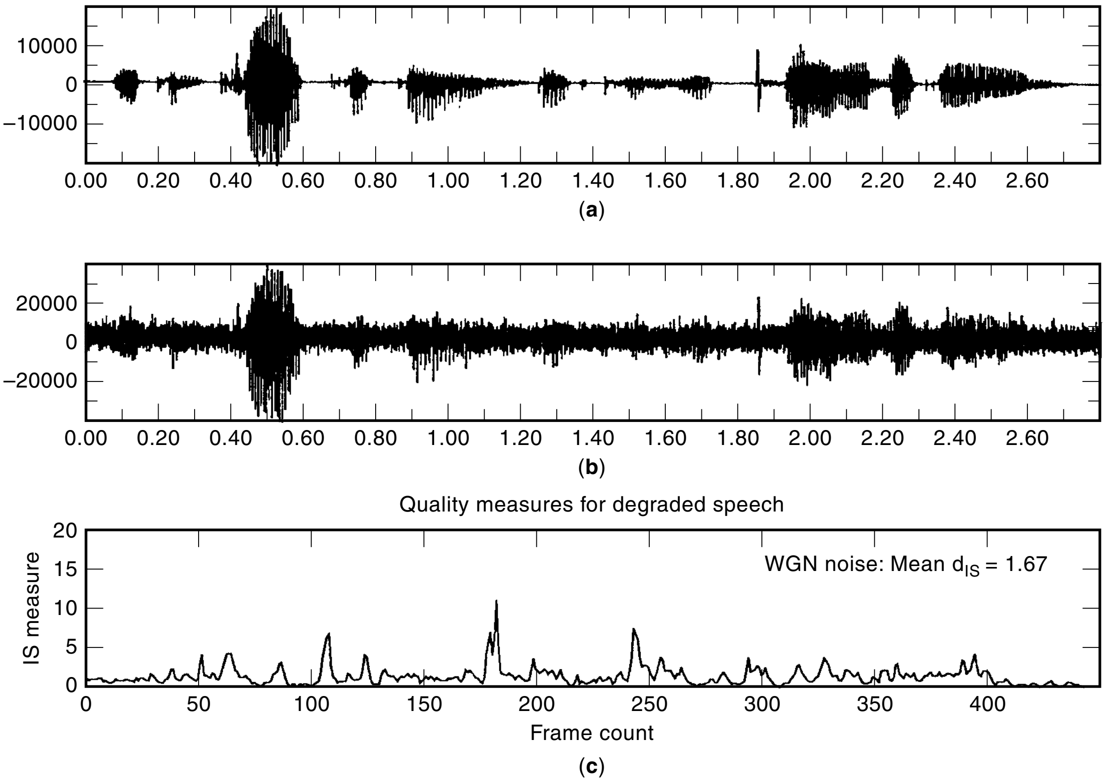

Figure 1. Nonuniform impact of background noise versus time for the sentence "You can talk the game, but can you play the game?" (a) Original speech waveform; (b) degraded speech waveform; (c) distortion versus time as measured by Itakura-Saito objective speech quality measure. Distortion is additive white Gaussian noise (WGN) at SNR = 5 dB.

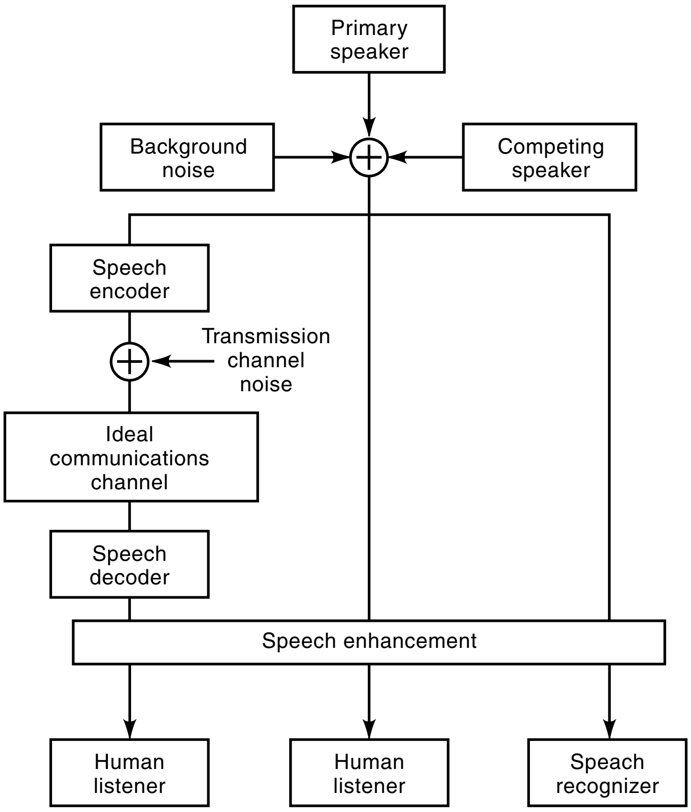

Figure 2. A general flow diagram of speech-enhancement applications and sources of distortion.

[Sheikhzadel, Brennan, and Sameti (98)]. It should also be noted that while it would seem obvious that speech enhancement would be useful for hearing-impaired listeners, other processing methods such as speech rate conversion may also be necessary for elderly hearing-impaired listeners [Nakamura et al. (10)].

Speech enhancement has also been employed as front-end processing stages for speech recognition [Juang (11), Hansen and Clements (12,13), Gong (14), Hansen and Arslan (15)] or back-end processing to improve speech-coding algorithms [Sen and Holmes (16)].

Speech-enhancement algorithms can also be used to improve communications or speech recognition in high-noise environments such as fighter aircraft cockpits [Harrison, Lim, and Singer (17); Darlington, Wheeler, and Powell (18); Hansen and Clements (19,20)].

Enhancement algorithms can be classified into two groups depending on whether a single-channel or dual-channel (or multichannel) approach is used. For *single-channel* applications, only a single microphone is available. Characterization of noise statistics must be performed during periods of silence between utterances, thus requiring a stationarity assumption of the background noise. In situations such as voice telephone or radio communications, only a single channel is available. In *dual-channel* algorithms, the acoustic sound waves arrive at each microphone at slightly different times (one normally is a delayed version of the other). Dual-channel enhancement assumes that a primary channel contains speech with additive noise and a second channel contains a noise signal reference. Typically some acoustic barrier must exist between microphones to ensure that no speech leaks into the noise reference channel. For multichannel methods, no acoustic barrier exists so the enhancement algorithm normally employs a beam-forming solution. In our discussion, we shall concentrate on methods that assume that (1) the noise distortion is additive, (2) the noise and speech signals are uncorrelated, and (3) that generally only one input channel is available.

The following four broad classes of enhancement techniques exist and will be considered in the following sections: (1) short-term spectral subtraction, (2) speech modeling using iterative Wiener-filtering methods, (3) adaptive noise canceling, and (4) methods based on fundamental frequency tracking.

In this article, we consider only a small subset of the possible topics of enhancement of speech degraded by noise. Specifically, we address the problem of speech degraded by additive noise as follows,

$$y(n) = s(n) + Gd(n) \tag{4}$$

where $s(n)$ is the original "clean" speech signal, $d(n)$ the degrading noise, $y(n)$ the degraded speech signal, and G a gain term that controls SNR.

SHORT-TERM SPECTRAL AMPLITUDE METHODS

Short-term spectral domain methods perform all their processing on the spectral magnitude. The enhancement procedure is performed over frames by obtaining the short-term magnitude and phase of the noisy speech spectrum, subtracting an estimated noise magnitude spectrum from the speech magnitude spectrum, and inverse transforming this spectral amplitude using the phase of the original degraded speech. Since background noise will degrade both the spectral magnitude and phase, it is reasonable to question the performance of a technique that does not address the noisy phase. It has been shown, however, that the human auditory system is relatively insensitive to phase for single-channel speech [Wang and Lim (21)]. For this reason, these methods seek to only enhance the noisy spectral magnitude.

Spectral Subtraction

Linear spectral subtraction is one technique based on direct estimation of the short-term spectral magnitude. It is assumed that the noise is short-term stationary, with second-order statistics estimated during silent frames (single channel) or from a reference channel (dual channel). The estimated noise power spectrum is subtracted from the transformed noisy input signal.

Let $s(n)$, $d(n)$, and $y(n)$ be sample data sets from three random processes as represented in Eq. (4). If we assume that $s(n)$ and $y(n)$ are short-term stationary, then this equation can be written in the spectrum domain as

$$\Gamma_y(\omega) = \Gamma_s(\omega) + \Gamma_d(\omega) \tag{5}$$

because $d(n)$ is an uncorrelated process. Here, the term *spectrum* is used to represent the frequency content of the each signal. Given $\Gamma_y(\omega)$ and an estimate of the noise spectrum $\hat{\Gamma}_d(\omega)$, it is possible to estimate the spectrum of the uncorrupted speech as

$$\hat{\Gamma}_s(\omega) = \Gamma_y(\omega) - \hat{\Gamma}_d(\omega) \tag{6}$$

While theoretically interesting, this analysis has little practical significance since we deal with real waveforms over short time frames. It does, however, suggest the essence of the spectral subtraction approach to noise elimination.

Let us now consider a more realistic approach and drop the stationarity assumptions on $\Gamma_y(\omega)$ and $\Gamma_s(\omega)$. Given a signal $y(n)$, the task is to estimate the corresponding speech $s(n)$. Recognizing that, at best, $s(n)$ will be "locally stationary" over short time ranges, we select a frame of $y(n)$, using a window of length N ending at time m, $f_y(n; m) = y(n)w(m - n)$. It follows that the selected frame can be expressed in terms of the underlying speech and noise frames as follows,

$$f_y(n;m) = f_s(n;m) + f_d(n;m) \tag{7}$$

By analogy to Eq. (6) it is possible to use the short-term power density spectra (stPDS) and estimate

$$\hat{\Gamma}_s(\omega;m) = \Gamma_y(\omega;m) - \hat{\Gamma}_d(\omega;m) \tag{8}$$

where we recall that the stPDS is defined as

$$\Gamma_y(\omega;m) = \frac{1}{N}\sum_{\eta=-\infty}^{\infty} r_y(\eta;m)e^{-j\omega\eta} \tag{9}$$

with $r_y(\eta;m)$ the short-term autocorrelation. Whereas the long-term PDS is part of a mathematical model that is only related to time waveforms in an abstract way, the same is not true of the stPDS of Eq. (8). In fact, the stPDS is related to the short-term discrete-time Fourier transform (stDTFT) in a simple way,

$$\Gamma_y(\omega;m) = \frac{S_y(\omega;m)S_y^*(\omega;m)}{N^2} = \frac{|S_y(\omega;m)|^2}{N^2} \tag{10}$$

For convenience, we assume that the factor $1/N$ is omitted from the definition in Eq. (9) so that we may write simply

$$\Gamma_y(\omega;m) = |S_y(\omega;m)|^2 \tag{11}$$

In effect, therefore, Eq. (8) offers a way to estimate the short-term magnitude spectrum of the speech, $|S_s(\omega; m)|$. Let us call the estimate $|\hat{S}_s(\omega; m)|$. While it may appear that an estimate of the noisy phase is also necessary, Wang and Lim (21) have determined that for all practical purposes, it is sufficient to use the *noisy* phase spectrum, $\varphi_y(\omega; m)$, as an estimate of the clean speech phase spectrum $\hat{\varphi}_s(\omega; m)$. Therefore, estimation of the frame of speech resulting from spectral subtraction is recovered from the stDFT estimate as

$$\begin{aligned}\hat{S}_s(\omega;m) &= |\hat{S}_s(\omega;m)|e^{j\hat{\varphi}_s(\omega;m)} \\ &= [\Gamma_y(\omega;m) - \hat{\Gamma}_d(\omega;m)]^{1/2}e^{j\varphi_y(\omega;m)}\end{aligned} \tag{12}$$

where $\Gamma_y(\omega; m)$ and $\varphi_y(\omega; m)$ are both obtained from the stDFT of the present noisy speech frame,

$$S_y(\omega;m) = |S_y(\omega;m)|e^{j\varphi_y(\omega;m)} = \Gamma_y^{1/2}(\omega;m)e^{j\varphi_y(\omega;m)} \tag{13}$$

and $\hat{\Gamma}_d(\omega; m)$ can be estimated using any frame of the signal in which speech is not present, or from a reference channel with noise only. This method is referred to as traditional spectral subtraction, and is illustrated in Fig. 3(a).

Variations to Spectral Substraction

Due to its simplicity in implementation, a number of variations for spectral subtraction are found in the literature. From a historical perspective, these are best illustrated by considering the generalized approach due to Weiss and Aschkenasy (22) given by

$$\hat{S}_s(\omega;m) = [|S_y(\omega;m)|^a - |\hat{S}_d(\omega;m)|^a]^{1/a}e^{j\varphi_y(\omega;m)} \tag{14}$$

where the power exponent a can be chosen to optimize performance. Regardless of the value of a, these techniques are often collectively referred to as *spectral subtraction,* but specific names are sometimes found in the literature. The case $a = 2$ as originally shown in Eq. (12) and used as the motivating case is sometimes referred to as *power spectral subtraction* because the noise removal is carried out by subtracting stPDS (squared short-term magnitude spectra). The name spectral subtraction is sometimes reserved for the case $a = 1$ in which removal is carried out by subtracting magnitude spectra. In fact, much of the basis for the ideas above are originally found in papers by Boll (23,24), which employs the $a = 1$ estimator, and by McAulay and Lampass (29), who derived the approach as a two-state power subtraction approach across individual spectral lines. Techniques using Eq. (14) directly with other values of a are sometimes called *generalized spectral subtraction* [Berouti, Schwartz, and Makhoul (25)].

Other variations exist in which the spectral subtraction is actually implemented in the time domain. A time-domain approach corresponding to Eq. (14) with $a = 2$ is called *correlation subtraction.* When $a = 2$, the magnitude spectral portion of the computation is essentially equivalent to estimating (the square root of)

$$\hat{\Gamma}_s(\omega;m) = \Gamma_y(\omega;m) - \hat{\Gamma}_d(\omega;m) \tag{15}$$

or, equivalently,

$$|\hat{S}_s(\omega;m)|^2 = |\Gamma_y(\omega;m)|^2 - |\hat{\Gamma}_d(\omega;m)|^2 \tag{16}$$

Since the short-term inverse discrete-time Fourier transform (stIDTFT) is a linear operation it follows immediately that

$$\hat{r}_s(\eta;m) = r_y(\eta;m) - \hat{r}_d(\eta;m) \tag{17}$$

where $\hat{r}_d(\eta; m)$ is an estimate of the short-term autocorrelation of the noise process. The time-domain approach can also be used for values of a, resulting in the name generalized correlation subtraction as shown in Fig. 3(b). Generalized spectral subtraction is shown in Fig. 3(c). From a historical perspective, we point out that the INTEL system of Weiss, Aschkenasy, and Parsons (26), which was the first reported spectral subtraction technique is based on correlation subtraction. A generalized cepstral processing version of the spectral estimator in Eq. (14) was part of an enhanced version of INTEL [Weiss and Aschkenasy (22)], which has been distributed and used extensively under the name speech en-

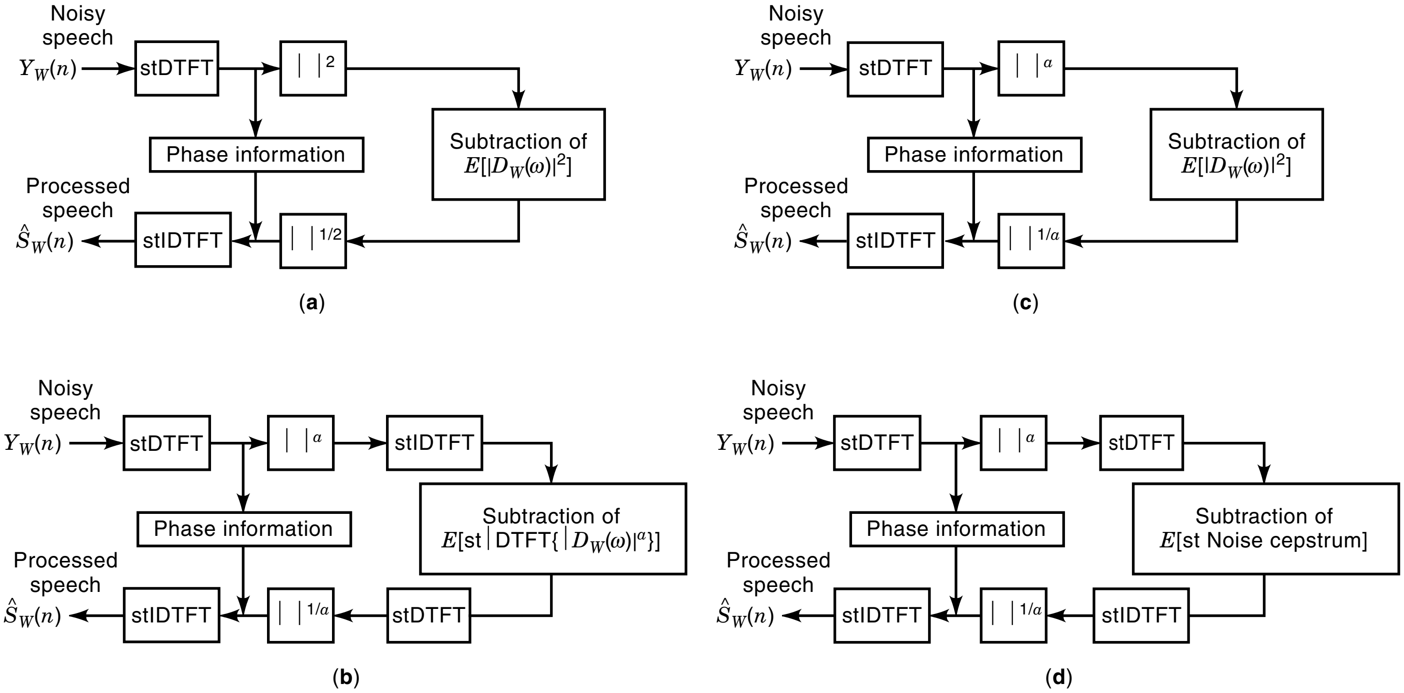

Figure 3. Variations in power spectral subtraction. (a) Original spectral subtraction; (b) generalized spectral subtraction; (c) generalized correlation subtraction; (d) improved INTEL correlation subtraction method. Here, stDTFT refers to the short-term discrete-time Fourier transform, and stIDTFT stands for the inverse of stDTFT.

hancement unit (SEU) by the U.S. Air Force (Rome Laboratory). Cepstral processing here refers to the inverse Fourier transform of the log-magnitude spectrum of the input speech signal. The systems in Fig. 3 summarize four variations of spectrum subtraction: (magnitude) spectral subtraction, generalized spectral subtraction, generalized correlation subtraction, and generalized cepstral processing.

Negative Spectral Components and Further Enhancements

From Eq. (12) it is observed that the estimated speech magnitude spectrum is not guaranteed to be positive. Different systems remedy this by performing half-wave rectification or full-wave rectification, or using a weighted difference coefficient. Most techniques use half-wave rectification (i.e., set negative portions to zero). Forcing negative spectral magnitude values to zero, however, introduces a "musical" tone artifact in the reconstructed speech. This anomaly represents the major limitation of spectral subtraction techniques.

The system proposed by Boll (23,24) attempts to reduce spectral error by applying magnitude averaging, which reduces spectral error by performing local averaging of the spectral magnitudes. Magnitude averaging works well if the time waveform is stationary. Unfortunately, the number of local averaging frames K is limited by the short-term stationarity assumption of speech. Therefore, only a few frames of data can be used in averaging (typically between 3 and 5 frames). In an evaluation using helicopter noise, Boll showed that spectral subtraction alone does not increase intelligibility but does increase quality, and that magnitude averaging does reduce the effects of musical tones caused by errors in accurate noise bias estimation.

In traditional spectral subtraction, a spectral estimate of the background noise is subtracted from the input noisy speech spectrum. One approach proposed by Ephraim and Malah (27) was to formulate a direct minimum mean-square error (MMSE) estimator for the short-time spectral amplitude component of speech in noise. This approach scales each noisy spectral magnitude value as follows, $|\hat{S}_i(\omega_k)| = H_i(\omega_k)|Y_i(\omega_k)|$, based on estimating the kth spectral magnitude,

$$H_i(\omega_k) = \left(\frac{\sqrt{\pi}}{2}\right)\left(\frac{\sqrt{v_k}}{\gamma_k}\right)\exp\left(-\frac{v_k}{2}\right) \left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] \tag{18}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ represent modified Bessel functions of the zero and first order, respectively, and v_k depends on both the *a priori* and *a posteriori* SNR of the kth spectral component. During enhancement, an estimate of the *a priori* SNR is obtained from previously enhanced spectral components. This estimator has been shown to be successful in suppressing background noise with few of the artifacts normally seen as musical tones in common spectral subtraction. A subsequent study by Cappe (28) explored the advantages of this estimator for restoration of musical recordings, and illustrated a smoothing procedure which could be used to obtain a more consistent estimate of the SNR. Subsequent improvements have been suggested to reduce these tone effects by Cappe (28).

Another approach which greatly influenced many ideas in spectral subtractions, was proposed by McAulay and Malpass (29), in which a particular spectral line is attenuated based on how much the speech plus noise power exceeds as estimate background noise floor. The noise at each frequency component is assumed to be Gaussian, resulting in a maximum likelihood estimate of $|S_s(\omega;\ m)|$. A further extension, also by McAulay and Malpass, is to scale the input frequency response $|S_y(\omega;\ m)|$ by the probability that speech is present in the input signal. Their reasoning is that if the probability of noise is high, it would be preferable to further reduce the signal estimate $|\hat{S}_s(\omega;\ m)|$. This contribution was particularly important, since it is essentially a two-state version of a hidden Markov model-based approach which was later formulated by Ephrain, Malah, and Juang (60,61). An extension to this approach was proposed by Hansen (30) in which a noise adaptive boundary detector was used to partition speech into voiced, transitional, and unvoiced speech sections to allow for a variable noise suppression based on the input speech class, followed by the application of morphological based spectral constraints to reduce frame-to-frame jitter of speech spectral characteristics. Performance was demonstrated for a variety of speech sound types (vowels, nasals, stops, glides, fricatives, etc.) over traditional spectral subtraction and noncausal Wiener-filtering techniques. Peterson and Boll (31) considered applying spectral subtraction in separate frequency bands, tuned to the loudness components perceived by the auditory system. Other extensions that are related to those presented here can be found in papers by Curtis and Niederjohn (32), Preuss (33), and Un and Choi (34). One method that considered a time-frequency partitioning was developed by Whipple (35) using a wavelet decomposition. Wavelet-based speech decomposition can be very effective for speech enhancement across changing phoneme content (i.e., better improvement for high frequency consonants such as fricatives or stops). Nonlinear spectral subtraction (NSS) by Lockwood and Boudy (36) takes into account the frequency-dependent SNR of colored noise. This algorithm reduces subtraction for spectral components of high SNR and increases subtraction for spectral components of low SNR. In addition, the noise model includes both an averaged noise spectrum and an overestimated noise spectrum. Further spectral subtraction extensions have also included a weighted subtraction term k in front of $|\hat{S}_d(\omega;\ m)|^a$ in Eq. (14), which is dependent on SNR [Berouti, Schwartz, and Makhoul (25), Arslan, McCree, and Viswanathan (37), George (38)] or an auditory masking threshold [Tsoukalas, Paraskevas, and Mourjopoulos (39); Virag (40)].

Dual-channel spectral subtraction has also been considered for the purposes of co-talker separation [Hanson, Wong, and Juang (41), Childers and Lee (42), Naylor and Boll (43), Morgan et al. (44)]. These methods normally require some *a priori* knowledge of the speaker characteristics (normally fundamental frequency contours) to assist in the enhancement process.

Spectral Subtraction Evaluation

Since a wide range of spectral subtraction methods exist, it will not be possible to consider the performance of most methods in different noise environments. Here, we briefly summarize two studies that considered three factors that greatly influence enhancement performance: (1) the enhancement domain, (2) the power factor term a (also related to the enhancement domain), and (3) processing of negative spectral components. In the first study, Lim and Oppenheim (45) evaluated the correlation subtraction method proposed by Weiss, Aschkenasy, and Parsons (26) called INTEL for wide-band random noise under varying values of a. Figure 4(a) shows intelligibility scores based on tests involving nonsense sentences. Results with wide-band noise show that intelligibility is not improved. It was also observed that processed speech with $a = 1$ or 0.5 sounded distinctly "less noisy" and of "higher quality" at relatively high SNR.

In a later study, Hansen and Clements (19,46) compared the performance of Boll's spectral subtraction method with that of traditional adaptive Wiener filtering (discussed in the next section). Evaluation was performed for both half- and full-wave rectification, employing one to five frames of magnitude averaging. Figure 4(b) summarizes some of these results using the Itakura-Saito measure from Eq. (2). It was shown that full-wave rectification resulted in improvement over a wider range of SNR; however, half-wave rectification had greater improvement over the restricted SNR band of 5 to 10 dB. In addition, magnitude averaging using frames that look ahead in time performed poorer than the corresponding equivalent looking back in time. For both rectification approaches, magnitude averaging provides improved levels of quality.

SPEECH MODELING AND WIENER FILTERING

A second speech enhancement area involves methods based on short-term Wiener filtering. Here, a frequency weighting for an optimum filter is first estimated from the noisy speech. The linear estimator of the uncorrupted speech $s(n)$, which minimizes a mean square error (MSE) criterion, is obtained by filtering $y(n)$ with a noncausal Wiener filter. This filter requires a priori knowledge of both speech and noise statistics, and therefore must also adapt to changing characteristics. In a single-channel framework, noise statistics must be obtained during silent frames. Also, since noise-free speech is not available, a priori statistics must be based upon $y(n)$, resulting in an iterative estimation scheme. One way to approximate the noncausal Wiener filter is to adapt the filter characteristics on a frame-by-frame basis by using the short-term PDS as follows:

$$H^{\dagger}(\omega; m) = \frac{\hat{\Gamma}_s(\omega; m)}{\hat{\Gamma}_s(\omega; m) + \hat{\Gamma}_d(\omega; m)} \tag{19}$$

noting that $\hat{\Gamma}_s(\omega;\ m)$ and $\hat{\Gamma}_d(\omega;\ m)$ are the estimated speech and noise spectra. Given the filter response $H^{\dagger}(\omega;\ m)$, the short-term speech spectrum is then obtained by filtering the noisy speech signal as

$$\hat{S}_s(\omega; m) = H^{\dagger}(\omega; m)S_y(\omega; m) \tag{20}$$

either in the time or frequency domain. Since $H^{\dagger}(\omega;\ m)$ has a zero phase spectrum, the output phase of the enhanced speech spectrum $\hat{S}_s(\omega;\ m)$ is simply the noisy phase from $S_y(\omega;\ m)$. Therefore, like spectral subtraction methods, adaptive Wiener filtering focuses its processing only in the spectral magnitude domain, but ends up attributing the same phase characteristic to the speech that is used in the spectral subtraction method.

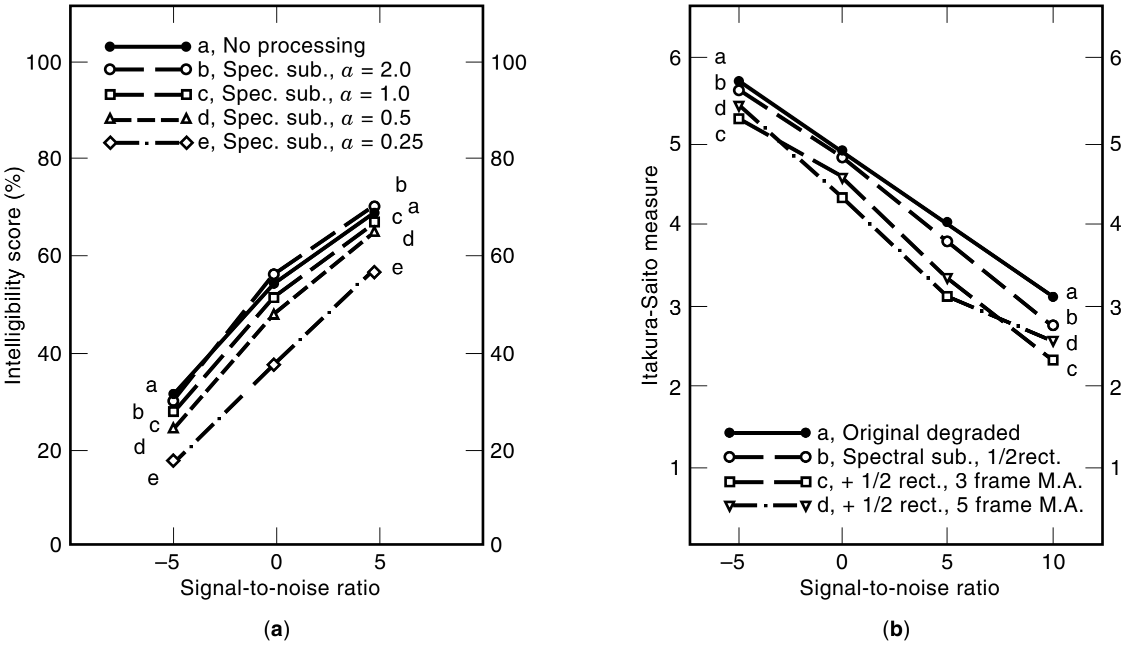

Figure 4. Performance evaluation of spectral subtraction methods: (a) intelligibility scores for different power exponent a terms [Lim and Oppenheim (45)]; (b) distortion as measured by Itakura–Saito objective speech quality measures for different levels of frame magnitude averaging [Hansen (93)]. Here, M.A. refers to magnitude averaging across frames as discussed in Boll (24).

One issue to address is the estimation of $\hat{\Gamma}_s(\omega; m)$ for the filter in Eq. (20). Since it is the speech in the frame that we are trying to estimate, it is unlikely that we would have an accurate estimate of its spectrum. One approach to the speech spectrum estimation problem is to use an iterative procedure in which an ith estimate of $\Gamma_s(\omega; m)$, say $\hat{\Gamma}_s(\omega; m, i)$ [or $|\hat{S}_s(\omega; m, i)|^2$]is used to obtain an $(i + 1)$st filter estimate, say $H^{\dagger}(\omega; m, i + 1)$. This approach to the estimation of speech parameters in an all-pole model assuming an additive white Gaussian noise distortion was investigated by Lim and Oppenheim (45) and later generalized for a colored-noise degradation by Hansen and Clements (19,20). The sequential method attempts to solve for the maximum *a posteriori* estimate of a speech waveform in additive white Gaussian noise with the requirement that the signal be the response from an all-pole process. The method was also generalized to include a pole-zero speech model (ARMA: auto-regressive moving average) by Musicus and Lim (47). Generalizations of the basic Wiener filtering in Eq. (20) have also been studied in other areas of signal processing. One approach for image restoration employs a noise scale term k and a power exponent a [Lim (48)] similar to that seen in generalized spectral subtraction [Eq. (14)].

Crucial to the success of noncausal Wiener filtering is the accuracy of the estimates of the all-pole parameters at each iteration. From these studies, it was shown that the estimation procedures that result in linear equations without background noise become nonlinear when noise is introduced. However, by using a suboptimal procedure, an iterative algorithm results that possesses the property that the estimation procedure is linear at each iteration.

Approaches Using Wiener Filtering

The basic sequential MAP (maximum *a posteriori*) estimation procedure by Lim and Oppenheim can be formulated in an alternate way. The estimate-maximize (EM) method was first introduced by Dempster, Laird, and Rubin (49) as a technique for obtaining maximum likelihood estimation from incomplete data. In the EM algorithm, the observations are considered “incomplete” with respect to some original set (which is considered “complete”). The algorithm iterates between estimating the sufficient statistics of the “complete” data given the observations and a current set of parameters (E step) and maximizing the likelihood of the complete data, using the estimated sufficient statistics (M step).

The EM approach is similar to the two-step MAP estimation procedure of Lim and Oppenheim; the main difference is that the error criterion is to maximize the expected log-likelihood function given observed or estimated speech data. Feder, Oppenheim, and Weinstein (50,51) formulated such a method for dual-channel noise cancellation applications where a controlled level of cross-talk was present. Their results showed improved performance over a traditional least-MSE estimation procedure.

Though traditional adaptive Wiener filtering is straightforward and useful from a mathematical point of view, there are several factors that make application difficult. Hansen and Clements (12,13,46) consider an alternative formulation based on iterative Wiener filtering augmented with speech-specific constraints in the spectral domain. This method was motivated by the following observations. First, although the sequential MAP estimation technique was shown to increase the joint likelihood of the speech waveform and all-pole parameters, a heuristic convergence criterion had to be employed. Second, as additional iterations were performed, individual formants of the speech consistently decreased in bandwidth and shifted in location. Third, frame-to-frame pole jitter was observed across time. Both of these effects contribute to unnatural sounding speech and illustrate a potential loss of stability in the iterative scheme. Only if the PDF of the unknown parameters is unimodal and the initial estimate for the speech parameters $\boldsymbol{a}_k$ is such that the local maximum equals the global maximum is the procedure equivalent to the joint MAP estimate of speech parameters $\hat{\boldsymbol{a}}_{i,n}$ and waveform $\hat{\boldsymbol{S}}_{i,n}$.

Hansen and Clements (13) proposed a variety of constrained iterative Wiener-filtering methods in order to improve parameter estimation, reduce frame-to-frame pole jitter across time, and provide a convenient and consistent terminating criterion. These methods were based on introducing spectral constraints between MAP estimation steps of LP (linear predictive) parameters and noise-free speech waveforms. The enhancement algorithms impose spectral constraints on all-pole parameters across time (interframe) and iterations (intraframe) that ensure that (1) the all-pole speech model is stable, (2) the model possesses speechlike characteristics (e.g., poles are not too close to the unit circle narrow bandwidths), and (3) the vocal tract characteristics do not vary wildly from frame to frame when speech is present. The most popular of these was Auto-LSP (see Fig. 5). In Auto-LSP, interframe constraints were applied on the line spectral pair (LSP) parameters to produce speech model pole movements across time, ensuring that formants lay along smooth tracks. Intraframe constraints were applied to the autocorrelation across iterations on a frame-by-frame basis. The imposition of these constraints helps in obtaining an optimal terminating iteration across all speech classes and improves speech quality by reducing the effects of traditional Wiener-filtering anomalies. As Fig. 5 illustrates, constraints are applied to the speech model parameters, and a constrained Wiener filter is constructed to obtain the next estimate of the speech waveform. Next, we consider quality improvement using this approach.

It should be noted that a comparison of speech enhancement algorithms can only be accomplished if evaluation conditions are equivalent (i.e., speech material, noise type and level, and quality or intelligibility testing methods). Figure 6(a) briefly summarizes IS objective speech quality results from a study by Hansen and Clements (13) that considered (1) noncausal (unconstrained) Wiener filtering [Lim and Oppenheim (45)], (2) spectral subtraction with magnitude averaging [Boll (24)], and (3) two inter- and intraframe spectral constrained Wiener-filtering methods. Quality measures for a theoretical limit were obtained by substituting the noise-free LP coefficients into the unconstrained Wiener filter, thereby requiring only one additional iteration to obtain the estimated speech signal. These results show that good quality improvement can be achieved with all three methods. Auto-LSP [plot *e* in Fig. 6(a)] did outperform noncausal Wiener filtering and spectral subtraction with magnitude averaging across all SNRs tested, though with higher computational requirements. However, Auto-LSP produces no musical tone artifacts, as is common in spectral subtraction methods. Later studies [Hansen and co-workers (12,15)] considered the performance in a variety of colored-noise environments.

Further Refinements to Iterative Filtering

Although all-pole modeling of speech has been used in many speech applications, it is known that some sounds are better modeled by a pole-zero system. Musicus and Lim (47,52) considered a generalized MAP estimation procedure based on a pole-zero model for speech. Essentially, the procedure requires MAP estimation of the predictor coefficients for both denominator and numerator polynomials, followed by MAP estimation of the noise-free speech through the use of an adaptive Wiener filter. Paliwal and Basu (53) considered a speech enhancement method based on Kalman filtering. A delayed-Kalman-filtering method was found to perform better than a traditional Wiener-filtering scheme. Another refinement proposed by Gibson and Koo (54) considers scalar and vector Kalman filters in an iterative framework in place of the adaptive Wiener filter for removal of colored noise. Further extensions based on aspects of the auditory system have been formulated by Nandkumar and Hansen (55–57), resulting in auditory constrained enhancement methods (ACE-I and ACE-II). The basic enhancement framework employs a dual-channel scenario using a two-step iterative Wiener-filtering algorithm. Constraints across broad speech sections and over iterations were then experimentally developed on a novel auditory representation derived by transforming the speech magnitude spectrum. The spectral transformations are based on modeling aspects of the human auditory process, which include critical band filtering, intensity to loudness conversion, and lateral inhibition. Objective speech quality results shown in Fig. 6(b) for speech degraded by slowly varying aircraft cockpit noise demonstrate the measurable improvement in IS measures when auditory processing constraints are employed.

Finally, other enhancement techniques based on speech modeling have employed vector quantization and a noisy based distance metric to determine a more suitable noise-free speech frame for enhancement [O'Shaughnessy (58); Gibson, Fisher, and Koo (59)]. Such methods require a training phase to characterize a speaker's production system. Another speaker-dependent enhancement approach by Ephraim, Malah, and Juang (60,61) employs a hidden Markov model (HMM) to characteize clean speech. The parameter set of the HMM is estimated using a clustering algorithm, followed by sequential estimation of the noise-free speech and HMM state sequences and mixture coefficients. The speech signal estimation process also results in a noncausal Wiener-filtering procedure. A later approach by Hansen and Arslan (62) incorporated HMM phoneme class partitioning to impose a phone-class-dependent terminating iteration. The method was shown to improve the consistency of the resulting enhanced speech over Auto-LSP constrained and unconstrained iterative Wiener filtering.

ADAPTIVE NOISE CANCELING

The general technique of adaptive noise canceling (ANC) has been applied successfully to a number of problems that include speech, electrocardiography, elimination of periodic interference, elimination of echoes on long-distance telephone transmission lines, and adaptive antenna theory. Initial work on ANC began in the 1960s and collectively refers to a class of adaptive enhancement algorithms that are based on the availability of a primary input source and a secondary reference source. While spectral subtraction and Wiener filtering can be generalized to operate in a dual-channel system, ANC usually requires a secondary reference channel. Initial studies on ANC can be traced to Widrow and co-workers at Stanford in 1965 and Kelly at AT&T Bell Laboratories. This work was later described by Widrow et al. (63). The adaptive line

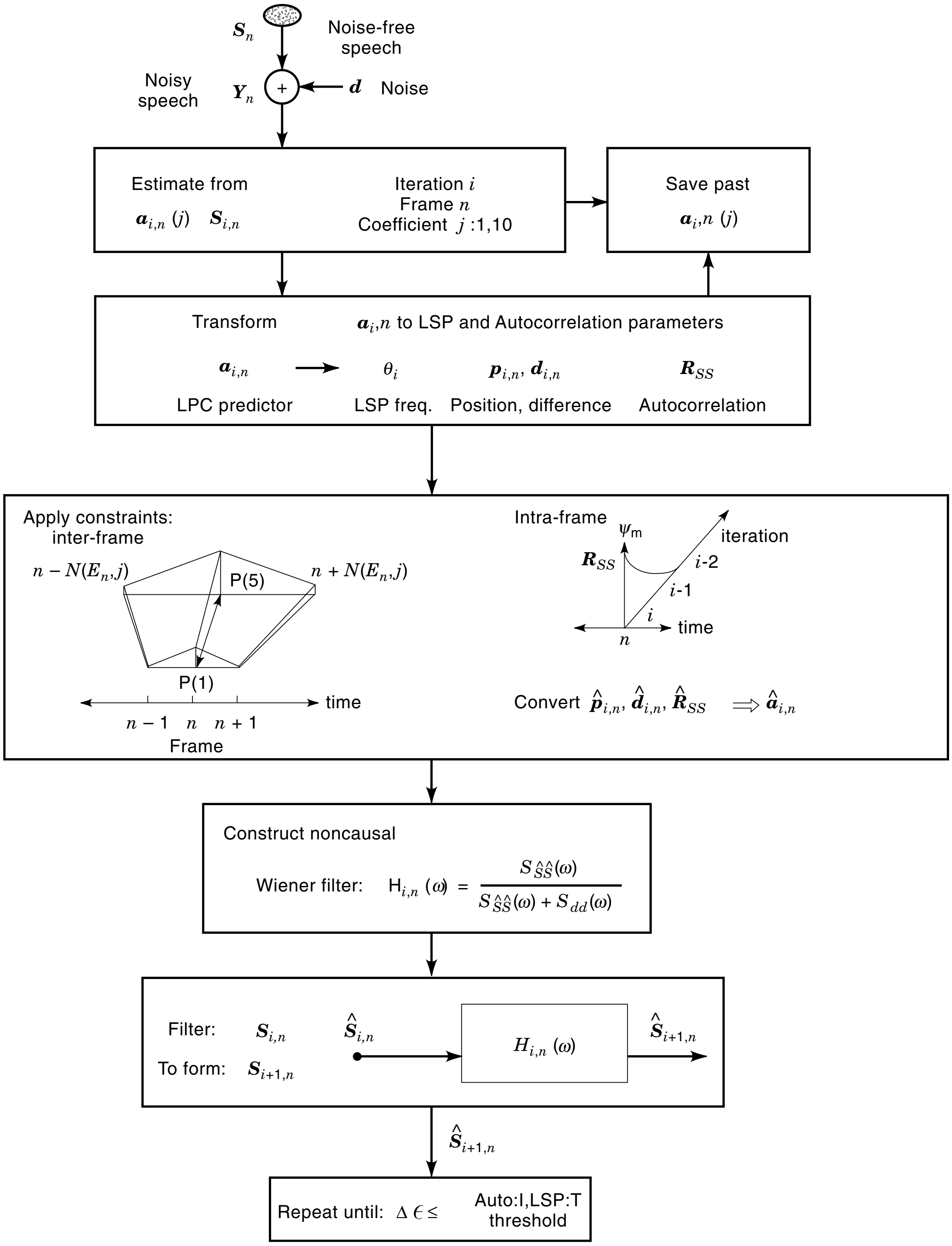

Figure 5. Speech enhancement based on all-pole modeling, noncausal Wiener filtering, and inter- and intraframe spectral constraints.

enhancer and its application as an adaptive detector were patented by McCool et al. in 1980 (64,65). Kelly, also in 1965, developed an adaptive filter for echo cancellation that uses the speech signal itself to adapt the filter. This work was later recognized by Sondhi (66). The echo canceler and its refinements by Sondhi are described in patents by Kelly and Logan (67) and Sondhi (68). Further details on the general area of adaptive filter theory can be found in texts by Haykin (69) and Messerschmitt (70).

ANC Based on the LMS Algorithm

The classical approach to dual-channel adaptive filtering, based on a least (or "minimum") mean square error (LMS)

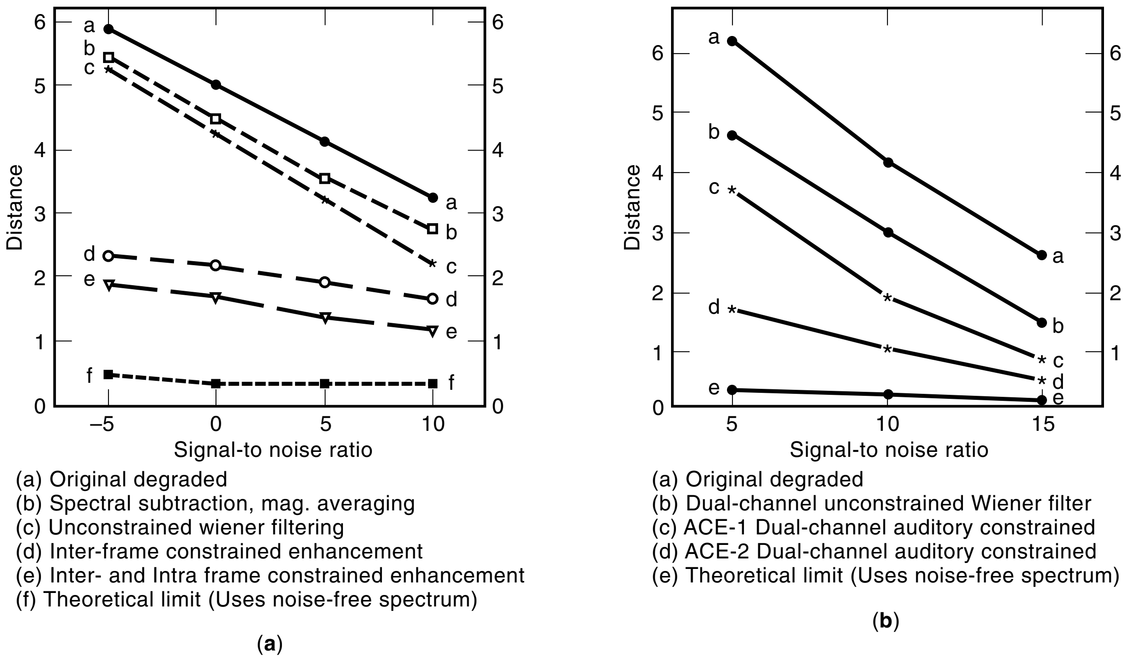

Figure 6. A comparison of constrained iterative speech-enhancement performance using the Itakura-Saito objective speech quality measure: (a) single-channel spectral subtraction and unconstrained Wiener filtering methods compared with interframe (FF-LSP : T) and inter- and intraframe (LSP : T, Auto : I) constrained enhancement methods [Hansen and Clements (13,93)]. (b) Dual-channel unconstrained Wiener filtering compared with auditory constrained interative speech enhancement methods (ACE-I, ACE-II) [Nandkumar and Hansen (55,57)].

criterion was first formulated by Widrow and co-workers (63,71). This technique has the major advantage of requiring no a priori knowledge of the noise signal. Figure 7 illustrates the LMS filter structure of a dual-channel adaptive noise canceler. All signals in this figure are assumed to be realizations of wide-sense-stationary stochastic processes with appropriate ergodicity properties so that we may use time waveforms in the analysis. The objective of the adaptive filter in Fig. 7 is to estimate the noise sequence $d_1(n)$ from $d_2(n)$ in order that the noise can be removed from $y(n)$. With this interpretation, the output of the noise canceler can be interpreted as an estimate, say $\hat{s}(n)$, of the uncorrupted speech $s(n)$. The filter is far infrared (FIR) with estimated tap weights, say $\hat{h}_i$, $i = 1, . . ., M$, so that

$$\hat{d}_1(n) = \sum_{i=1}^{M} h_i d_2(n - i + 1) \tag{21}$$

A natural optimization criterion is to minimize the mean-square error (MSE) between the sequences $d_1(n)$ and $\hat{d}_1(n)$. Unfortunately, the signal $d_1(n)$ is not measurable. However, it can be shown that attempting to estimate $d_1(n)$ using $d_2(n)$ with a least MSE criterion is equivalent to estimating $d_1(n)$ plus any signal that is orthogonal to $d_2(n)$ [meaning that $d_1(n)$ and $d_2(n)$ are orthogonal random processes]. Since we generally assume that the speech signal $s(n)$ is uncorrelated to the degrading noise signal $d_1(n)$ [and $d_2(n)$], we may attempt to estimate $y(n)$ from $d_2(n)$ and derive an identical filter to that which would be obtained for estimating $d_1(n)$. Since the only part of $y(n)$ that is correlated with $d_2(n)$ is $d_1(n)$, the best estimate of $y(n)$ will in fact be the best estimate of $d_1(n)$. It is interesting to note that for this interpretation the signal $\hat{s}(n)$ is interpreted as an *error* [call it $\epsilon(n)$] that is to be minimized in mean-square sense. Therefore, the ANC is sometimes described as having been designed by minimizing its output power (or energy in the short-term case).

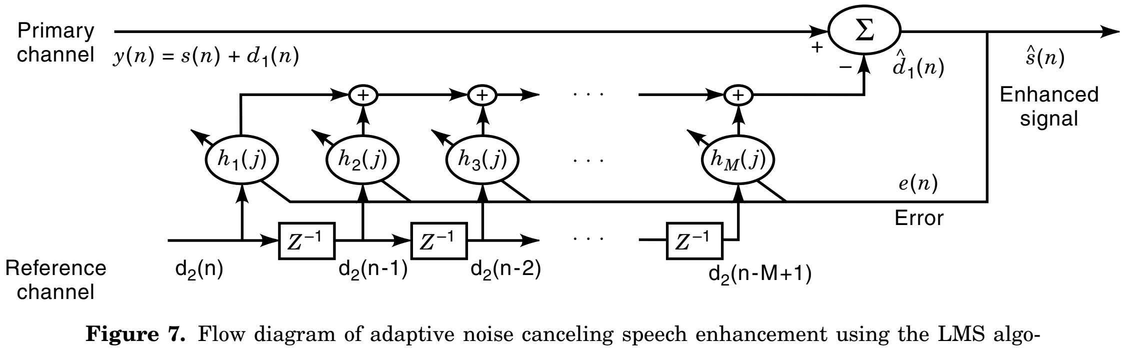

Figure 7. Flow diagram of adaptive noise canceling speech enhancement using the LMS algorithm.

With this alternative, but equivalent, optimization criterion, $\hat{h}$ is chosen such that

$$\begin{aligned}\hat{h} &= \arg\min_{h} \mathscr{E}\{[y(n) - \hat{y}(n)]^2\} \\ &= \arg\min_{h} \mathscr{E}\left\{\left(y(n) - \sum_{i=1}^{M} \hat{h}_i d_2(n-i+1)\right)^2\right\}\end{aligned} \tag{22}$$

or, in a matrix-vector notation,

$$\boldsymbol{R}_{d_2}\hat{h} = \boldsymbol{r}_{y,d_2} \tag{23}$$

The remaining issue is the solution of Eq. (23) for the filter tap weights $\hat{h}$. ANC designs have frequently employed the LMS algorithm, which has been demonstrated to be an effective and practical means for real-time approximation of the filter solution in this application [Widrow et al. (72,73)]. Let us denote the MSE for the general filter $\hat{h}$ by $\xi(\hat{h})$. The minimum MSE is therefore $\xi(\hat{h}^{\dagger})$. The error $\xi(\hat{h})$, when considered as a function of the weights, is frequently called an *error surface.* Now, by definition

$$\xi(\hat{h}) = \mathscr{E}\left\{\left(y(n) - \sum_{i=1}^{M} \hat{h}_i d_2(n-i+1)\right)^2\right\} \tag{24}$$

Since this expression is a quadratic in $\hat{h}$, there is a unique minimum that occurs at $\hat{h}^{\dagger}$. The LMS algorithm gradually moves toward the minimum by "slowly" moving against the gradient of the error surface. A solution is obtained by differentiating with respect to the entire weight vector at once,

$$\frac{\partial \xi(\hat{h})}{\partial \hat{h}} = \boldsymbol{r}_{y,d_2} - \boldsymbol{R}_{d_2}\hat{h} \tag{25}$$

The resulting recursion is

$$\hat{h}^{n} = \hat{h}^{n-1} - \Delta^{n}\hat{g}^{n} \tag{26}$$

where $\hat{g}^n$ indicates the estimated gradient associated with time n. In practice, a fixed step size, $\Delta^n = \Delta$, is often used for ease of implementation and to allow for adaptation of the estimate over time as the dynamics of the signal change. Since background noise characteristics can also change across time, the FIR filter coefficients are typically adapted as function of time [note that in Fig. 7, the FIR tap weights $\hat{h}_i(j)$, $i = 1, \ldots, M$ have a common time index j]. This simple algorithm was first proposed by Widrow and Hoff (71) and is now widely known as the LMS algorithm. The convergence, stability, and other properties of LMS have been studied extensively and are discussed in great detail by [Widrow and Stearns (74)]. It has been shown using long-term analysis [Widrow et al. (63)] that starting with an arbitrary initial weight vector, the algorithm will converge in the mean and remain stable as long as the following condition on the step size parameter Δ is satisfied,

$$0 < \Delta < \frac{1}{\lambda_{\max}} \tag{27}$$

where $\lambda_{\max}$ refers to the largest eigenvalue of the autocorrelation matrix of the reference channel R_{d_2}.

Experimental Applications of ANC

One of the advantages of dual-channel adaptive noise canceling is that speech with either stationary or nonstationary colored noise can be processed. In general, the two microphones are required to be sufficiently separated in space, or contain an acoustic barrier between the two microphones to achieve noise cancellation. It should be noted also, that ANC cannot remove white noise, since samples of white noise are uncorrelated, meaning that the adaptive filter could not predict $d_1(n)$ from $d_2(n)$.

One of the earlier dual-channel evaluations of ANC for speech was conducted by Boll and Pulsipher (75). Two adaptive algorithms were investigated: the LMS approach of Widrow et al. (76) and the gradient lattice approach of Griffiths (77), which employs a lattice filter framework rather than other methods that use tap-delay lines (FIR filters). The lattice filter approach was suggested since the typical FIR adaptive filter necessary to estimate the input noise characteristics required 1500 tap weights. Such large filter lengths result in misadjustment and are therefore an important design criterion since large misadjustment leads to pronounced echo in the resulting speech signal. Fortunately, the echo can be reduced by reducing the adaptation step size used in updating the filter weights, but this increases the settling time of the adaptive filter. Boll and Pulsipher also suggested a frequency-domain LMS adaptive filter, which can result in a substantial savings in computation, though additional throughput processing delay is experienced using this method. Figure 8(a) summarizes background noise suppression versus time for the three ANC methods. The major points from this study suggest that LMS or gradient-lattice-based ANC can provide noise suppression in the time domain but that a large tap delay filter is needed. While all three dual-channel methods provide measurable noise suppression, the settling time is less for the short-time Fourier transform approach (ANC-STFT) than for the lattice approach (ANC-lattice). This comes at the expense of additional processing delay between input and output speech samples.

An important application of dual-channel adaptive noise cancellation in which large microphone spacing is not an issue is in aircraft cockpit environments. In this case, the pilot's oxygen facemask serves as acoustic barrier between the two sensors, thereby ensuring that the SNR of the primary sensor be much greater than the SNR of the reference sensor, while permitting close sensor spacing. Many aspects of the cockpit noise problem have been studied. The interested reader is referred to papers by Harrison, Lim, and Singer (17,78); Darlington, Wheeler, and Powell (18); Powell, Darlington, and Wheeler (79); and Rodriguez, Lim, and Singer (80). In the study by Harrison, Lim, and Singer, ANC was employed in a fighter cockpit environment. Results summarized in Fig. 8(b) show that the average SNR improvement ranged from +11.6 to +11.2 dB for input SNRs of 3 to 10 dB respectively. They concluded that a filter length of 100 taps was sufficient to achieve +11.4 dB improvement in SNR, and that increasing the filter tap length to 1000 with an exact least-squares method only improved the SNR by 3 dB. They point out that since there is typically more than one noise source in real fighter cockpit environments that may also be distributed over a region, performance of ANC generally degrades when the noise source is not localized to one location.

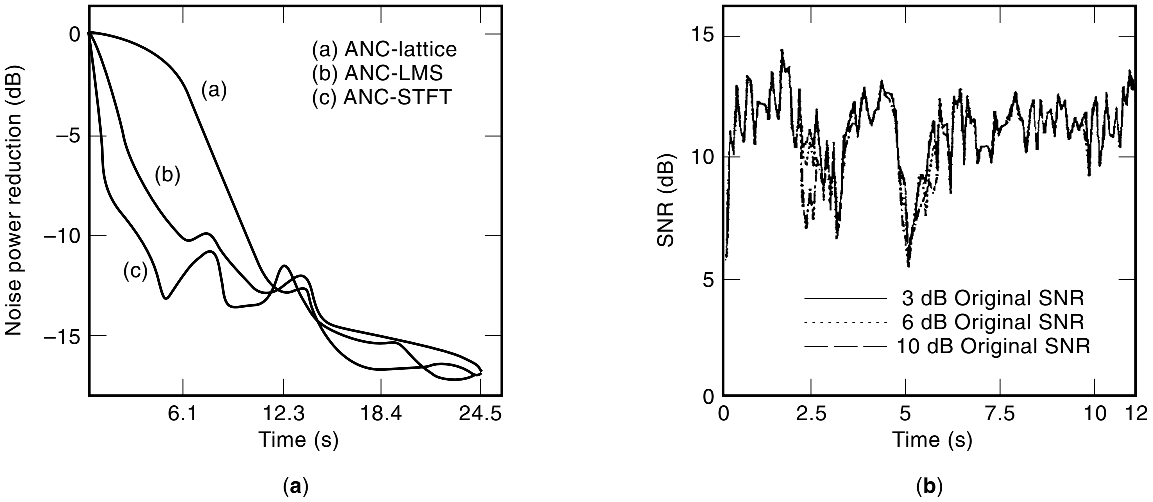

Figure 8. Enhancement performance of adaptive noise canceling: (a) a comparison of three different time- or frequency-domain implementations [Boll (94)]; (b) noise suppression versus time for the ANC-LMS method [Harrison, Lim, and Singer (17,78)].

One final area which is related in some sense to ANC is *Active* Noise Cancellation (AcNC). Here, AcNC is a noise suppression approach in which a secondary noise source is introduced that destructively interferes with the unwanted noise. These systems typically rely on multiple sensors to measure the unwanted noise field and produce an output acoustic signal which is out-of-phase with noise at the primary microphone. Since the artificially produced noise is out-of-phase, it will add destructively with the noise in the primary microphone. Several recent studies have considered single-sensor [Oppenheim, et al. (95)] and dual-sensor [Zangi (96)] approaches. The general area of active noise cancellation has been applied within car interiors, aircraft engines, and other sound enclosures. An excellent review can be found in Elliott and Nelson (97).

METHODS BASED ON FUNDAMENTAL FREQUENCY TRACKING

In the field of speech enhancement, there are a number of techniques that are based on tracking the fundamental frequency contour. Such approaches include single-channel ANC, adaptive comb filtering, and enhancement based on harmonic selection or scaling. These techniques capitalize on the property that waveforms during voiced passages are periodic. This periodicity ideally results in a line spectrum in the frequency domain. Any spectral components between these lines represent noise which can be reduced. One useful application for these techniques has been the competing speaker problem, in which the enhancement takes advantage of differences in fundamental frequency contours.

Single-Channel ANC

It has been shown that ANC employing the LMS algorithm requires no a priori knowledge of the noise signal. Generally speaking, ANC can only be employed when a second channel is available. Suppose however, we could simulate a reference using data from the primary channel. Sambur (81) proposed such an approach where instead of canceling noise in the primary channel, the speech signal is canceled. While it may be difficult to form a noise reference channel, it is not difficult to obtain a speech reference channel for some classes of speech. Due to the quasiperiodic nature of speech during voiced sections, a reference signal can be formed by delaying the primary data by one or two pitch periods. This reference signal can then be used in the LMS adaptive algorithm, for which the criterion is to form the minimum MSE estimate of the clean speech signal. Let the reference signal $y_2(n)$ be a delayed version of the primary signal, $y_2(n) = y_1(n - T_0)$, where T_0 represents one pitch-period delay. Then under ideal periodic speech conditions, we have

$$y_2(n) = s(n - T_0) + d(n - T_0) = s(n) + d(n - T_0) \qquad (28)$$

The delayed speech signal $s(n - T_0)$ will be highly correlated with the original speech $s(n)$, while the delayed $d(n - T_0)$ and original $d(n)$ noise signals will have low correlation with the speech signal. The flow diagram shown in Fig. 9 represents such an approach proposed by Sambur (82). Although the primary output is the "enhanced" noise signal $\hat{d}(n)$, an enhanced speech signal output $\hat{s}(n)$, is also available. The LMS adaptive filter produces the following output,

$$\hat{s}(n) = \sum_{i=1}^{M} \hat{h}_i y_1(n - T_0 - i + 1) \qquad (29)$$

where $\hat{h}_i$, = 1, . . ., M are the FIR filter weights, obtained in a manner similar to that discussed in section entitled "ANC based on the LMS Algorithm." Since this method exploits input speech periodicity, in principle it should only be applied for voiced speech. Therefore, noisy unvoiced speech could either be passed through the system unprocessed, or the LMS fitler response could be held constant and allowed to process the unvoiced speech.

Sambur (81,82) evaluated this approach and showed improved quality for additive white noise in the SNR range 0–10 dB. It was observed that the more severe the noise, the

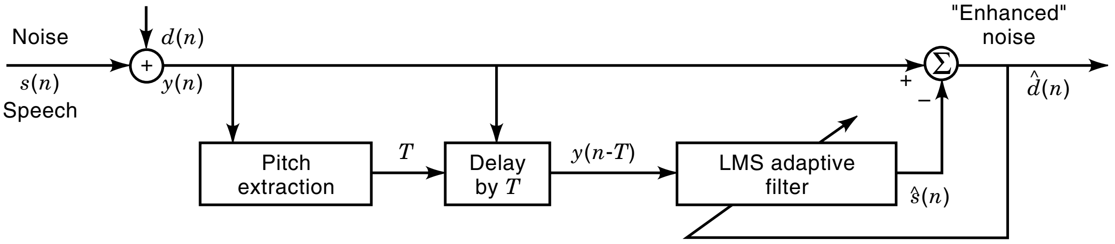

Figure 9. Single-channel adaptive noise cancellation based on fundamental frequency tracking.

more dramatic the improvement in SNR. Increased FIR filter length also improved performance. Listeners concluded that the speech was more pleasant to listen to and "appeared" to have more intelligibility (though no formal intelligibility tests were performed). Performance in the presence of speech coder quantization noise from a variable rate delta modulation system was also determined. The LMS adaptive filter removed some of the "granular" quality of the quantized speech. ANC was able to remove the granular noise since it is signal-independent and broadband, but leaves slope overload noise unaffected, since it is signal dependent. Later studies have also shown that spectral subtraction can be useful as a front-end processor for voice coding applications.

One of the main limitations of single-channel ANC is the requirement of accurate pitch estimation. Modifications to Sambur's method have been proposed by Varner, Miller, and Eger (83) in which a reference signal is obtained through the use of a low-order DPCM adaptive predictor (see the related article entitled "SPEECH CODING"). Kim and Un (84) also attempt to remove the pitch estimator by using both forward and backward adaptive filters.

Adaptive Comb Filtering

Corrupting noise can take many forms. In some applications speech is degraded by an underlying process which is periodic, resulting in a noise spectrum which also possesses periodic structure. Two methods are available for reducing such noise, which include adaptive comb filtering (ACF) [Lim and Oppenheim (85)] and time domain harmonic scaling (TDHS) [Malah and Cox, 1979 (86)].

ACF is similar in its basic assumptions to single-channel LMS-based ANC. If the noise is nonperiodic, its energy will be distributed throughout the spectrum. The basic process of comb filtering is to build a filter that passes the harmonics of speech, while rejecting noise frequency components between the harmonics. A typical block diagram for an adaptive comb filter is shown in Fig. 10. The comb filter has large values at the specified fundamental frequency F_0 and its harmonics, and low values between. The filter is usually implemented in the time domain as

$$\hat{s}(n) = \sum_{i=-L}^{L} c(i)y(n - iT_0) \tag{30}$$

where $c(i)$ are the $2L + 1$ comb filter coefficients, T_0 is the fundamental period in samples, and L is a small constant (typically 1–6) that represents the number of pitch periods used forward and backward in time for the filtering process.

Since only voiced speech can be enhanced, alternative processing is needed for unvoiced speech or silence. One approach is to pass the unvoiced speech through the filter unprocessed [i.e., set $c(k) = 0$ for all $k \neq 0$]. For this case, a scaling term (typically in the range 0.3 to 0.6) is necessary because applying an adaptive comb filter to voiced sounds reduces the noise energy present. This ensures a proper balance in the resulting signal strength between voiced and unvoiced sections. A second processing approach for unvoiced speech is to maintain a constant set of filter coefficients from the last voiced speech frame and process the unvoiced sounds as if they were voiced. This technique has not been as successful as the first.

In general, since F_0 typically changes within an analysis window, this can be addressed by including timing adjustments ζ_i for $y(n - iT_0)$ in Eq. (30) to align adjacent pitch periods. This results in a comb filter that is adaptive. It should also be noted that it is desirable to include as many periods as possible, since the number L is inversely proportional to the bandwidth of each tooth in the comb filter. Larger values of L therefore produce more narrow harmonics for the filter and therefore allow for furthernoise removal.

Malah and Cox (87) proposed a generalized comb-filtering technique that applies a time-varying weight to each pitch period. The generalized comb filter was shown to reduce

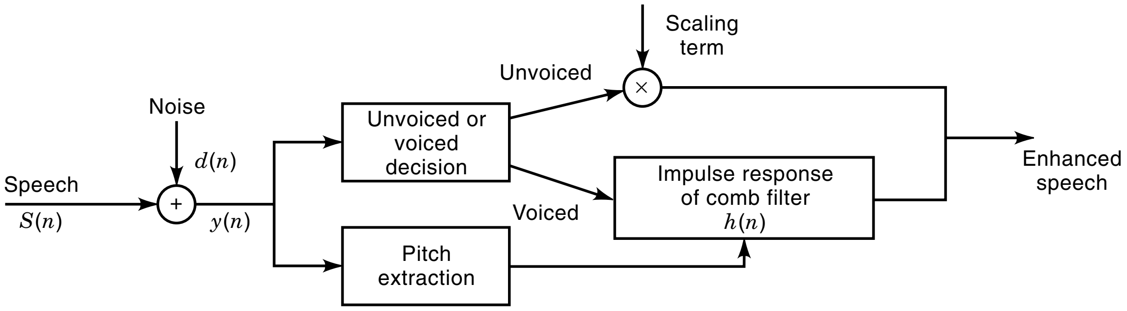

Figure 10. Flow diagram of the adaptive comb-filtering algorithm.

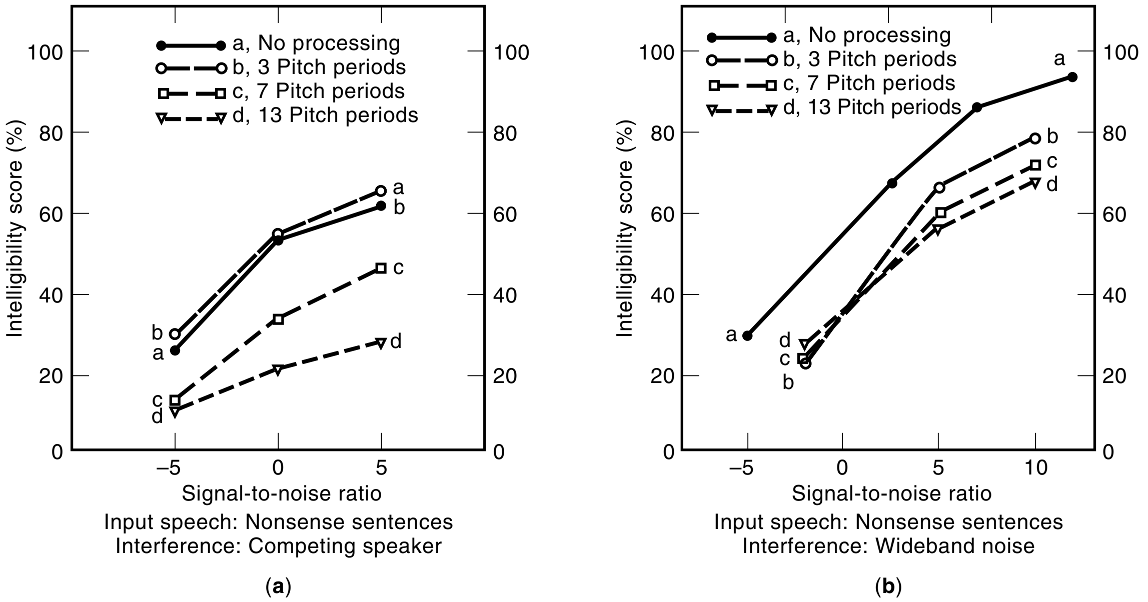

Figure 11. Intelligibility results for adaptive comb filtering for the problems of (a) competing speaker [Perlmutter et al. (68)] and (b) wide-band random noise [Lim and Oppenheim (45)].

frame-rate noise for an adaptive transform coder [Cox and Malah (88)]. Perlmutter et al. (89), using a competing speaker [Fig. 11(a)], and Lim, Oppenheim, and Braida (90), using wide-band random noise [Fig. 11(b)], evaluated this adaptive technique for varying filter length. Using pitch information obtained from noise-free speech, decreases in intelligibility were usually observed for various SNR. Both studies mention that processed speech sounded "less noisy" due to the system's ability to increase the local SNR (though no quality test were performed).

Time-Domain Harmonic Scaling

An alternative to frequency-domain harmonic selection is time-domain harmonic scaling (TDHS). TDHS can be viewed as a time-domain technique that requires pitch synchronous block decimation and interpolation. The difference between TDHS noise reduction and adaptive comb filtering is that TDHS moves the noise into the gaps under each pitch harmonic (i.e., masking the background noise), while comb filtering seeks to filter out the noise in the gaps between harmonics.

TDHS was originally proposed by Malah and Cox (86) for use in perceptually reducing periodicity structured noise in speech. In a later study, Cox and Malah (88) proposed a hybrid system that uses both ACF and TDHS. An additional benefit of their system is time-scale reduction of input speech for waveform coding and isolated word recognition. Due to its time-domain implementation, the choice of an appropriate window will greatly influence noise cancellation performance.

FUTURE DIRECTIONS FOR SPEECH ENHANCEMENT

In this article, we have considered a variety of approaches for speech enhancement. Due to the wide number of applications and assumptions concerning interference and available input channels, almost an unlimited number of enhancement systems could have been considered.

Four general classes of speech-enhancement algorithms were considered: (1) short term spectral amplitude methods, (2) speech modeling and iterative Wiener filtering, (3) adaptive noise canceling, and (4) fundamental-frequency-tracking methods. Application-specific speech enhancement has also played an important role in the development of algorithms in these areas. In the four areas, approaches based on spectral subtraction (short-term spectral amplitude) have by far been the most popular. The main reason for this is their ease in implementation on computer workstations and real-time DSP platforms. However, the residual musical tone artifacts that persist in the resulting enhanced speech have been the focus of much of the subsequent research since the studies by McAulay and Malpass (29), Boll (24), and Weiss and Aschkenasy (22,26). Methods based on iterative Wiener filtering have been shown to be effective in suppressing white and colored background noise beyond that in some forms of spectral subtraction. However, this comes at the expense of a measurable increase in computational requirements and system complexity. Methods based on adaptive noise canceling have been popular for use in actual noise environments if a second reference microphone is available. Since a pilot's oxygen facemask serves as a natural acoustic barrier between the two microphones, ANC has been shown to be effective in aircraft cockpit applications. ANC continues to be a strong reliable approach for a broad range of background noise environments. The main reason for this is that the FIR adaptive filter is well suited for implementation on real-time DSP processing platforms, and ANC makes no assumptions about the background noise statistics. However, it has been shown that ANC works better when the interfering noise source is localized to one location [Harrison, Lim, and Singer (17)]. Methods based on fundamental frequency tracking have not been popular for broadband noise distortion. This is because these methods can only improve the quality of voiced speech, and for such environments the degradation in consonant sections contributes significantly to losses in intelligibility. However, fundamental-frequency-tracking methods such as comb filtering and harmonic selection have proved to be useful for the competing speaker problem. In most demonstrations, pitch information is usually assumed, and the speakers are normally assumed to have dramatically different pitch structure (i.e., a male versus a female speaker).

In concluding, it should be emphasized that many enhancement systems improve the ratio of speech to noise and therefore improve quality. This is especially true if the context of the test material is known to the listener, so that intelligibility issues are not of concern. However, the majority of speech-enhancement algorithms actually reduce intelligibility, and those that do not generally degrade the quality. This balance between quality and intelligibility suggests that considerable work remains to be done in the area of speech enhancement.

Furthermore, as we have seen here, several enhancement algorithms have been shown to improve a mathematical criterion. Though attractive in a mathematical sense, most choices with respect to error criteria are not well correlated with auditory perception. This can be attributed to in part by the limitations that exist in present-day speech-modeling techniques and lack of effective evaluation methods. The use of subjective and objective quality measures should provide a firm foundation from which to compare enhancement performance. Evaluation of an enhancement algorithm will depend on its ultimate application.

Future directions in the field of speech enhancement will focus more on applications in which environmental interference (i.e., noise, channel, microphone, voice coding methods, or reverberation) impact the quality and intelligibility of speech. As a number of recent studies have shown, the trend is to incorporate further knowledge of auditory processing in conjunction with signal-processing concepts and criteria. Some of these have included lateral neural inhibition [Cheng and O'Shaughnessy (91)], auditory constrained iterative Wiener filtering [Nandkumar and Hansen (55–57)], and concepts on auditory noise masking thresholds [Tsoukalas and coworkers (39,92); Virag (40)]. Such advances will lead to improvements in voice communications and wireless telephony, digital hearing aids, and better man-machine interfaces for speech recognition.

We have identified quality, intelligibility, and reducing listener fatigue as three goals for enhancement. If speech enhancement is needed for continuous speech communications (air traffic control communications), annoying artifacts or uneven quality improvement (improvement in vowels, distortion in fricatives) may result in higher levels of listener fatigue. Therefore, a system designer might choose an algorithm with lower overall quality improvement if the resulting speech quality is more consistent across all speech classes.

As these concluding comments suggest, speech enhancement continues to be an important research area for improving communications between speaker and listener, speaker and vocoder, or speaker and speech recognizer.

BIBLIOGRAPHY

1. J. Lim (ed.), *Speech Enhancement.* Englewood Cliffs, NJ: Prentice-Hall, 1983.
2. J. Deller, J. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals,* MacMillan Series. New York, NY: Prentice-Hall, 1993.
3. D. O'Shaughnessy, Enhancing speech degraded by additive noise or interfering speakers. *IEEE Commun. Mag.,* **27** (2): 46–52, 1989.
4. Y. Ephraim, Statistical-model-based speech enhancement systems. *Proc. IEEE,* **80**: 1526–1555, 1992.
5. S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
6. P. L. Chu and D. G. Messerschmitt, A weighted Itakura-Saito spectral distance measure. *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-30**: 545–560, 1982.
7. F. Itakura, Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust. Speech Signal Process.,* **23**: 67–72, 1975.
8. J. M. Kates and M. R. Weiss, A comparison of hearing-aid array-processing techniques. *J. Acoust. Soc. Am.,* **99** (5): 3138–3148, 1996.
9. N. A. Whitmal, J. C. Rutledge, and J. Cohen, Reducing correlated noise in digital hearing aids. *IEEE Eng. Med. Biol. Mag.,* **15** (5): 88–96, 1996.
10. A. Nakamura, N. Seiyama, R. Ikezawa, T. Takagi, and E. Miyasaka, Real time speech rate converting system for elderly people. In *Proc. 1994 IEEE ICASSP,* Vol. 2, pp. 225–228, April 1994.
11. B. H. Juang, Speech recognition in adverse environments. *Comput. Speech Lang.* **5**: 275–294, 1991.
12. J. H. L. Hansen and M. A. Clements, Constrained iterative speech enhancement with application to automatic speech recognition. In *Proc. 1988 IEEE ICASSP,* pp. 561–564, April 1988.
13. J. H. L. Hansen and M. Clements, Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. Signal Process.,* **39**: 795–805, 1991.
14. Y. Gong, Speech recognition in noisy environments: A survey. *Speech Commun.* **16**: 261–291, 1995.
15. J. H. L. Hansen and L. M. Arslan, Robust feature estimation and objective quality assessment for noisy speech recognition using the credit card corpus. *IEEE Trans. Speech Audio Process.,* **3**: 169–184, 1995.
16. D. Sen and W. H. Holmes, Perceptual enhancement of CELP speech coders. *Proc. 1994 IEEE ICASSP,* Vol. 2, pp. 105–110, April 1994.
17. W. A. Harrison, J. S. Lim, and E. Singer, Adaptive noise cancellation in a fighter cockpit environment. In *Proc. 1984 IEEE ICASSP,* 18A.4.1–4, March 1984.
18. P. Darlington, P. D. Wheeler, and G. A. Powell, Adaptive noise reduction in aircraft communication systems. In *Proc. 1985 IEEE ICASSP,* pp. 716–719, March 1985.
19. J. H. L. Hansen and M. A. Clements, Enhancement of speech degraded by non-white additive noise. Final Technical Report submitted to Lockheed Co., No. DSPL-85-6, Georgia Institute of Technology, Atlanta, August 1985.
20. J. H. L. Hansen and M. A. Clements, Objective quality measures applied to enhanced speech. In *Proc. Acoust. Soc. Amer.,* 110th Meeting, C11, Nashville, Tenn., Nov. 1985.
21. D. L. Wang and J. S. Lim, The unimportance of phase in speech enhancement, *IEEE Trans. Acoust. Speech Signal Process.* **30**: 679–681, 1982.
22. M. R. Weiss and E. Aschkenasy, Computerized audio processor. Final Report, Rome Air Development Center, RADC-TR-83-109, May 1983.
23. S. F. Boll, Suppression of noise in speech using the SABER method. In *Proc. 1978 IEEE ICASSP,* pp. 606–609, April 1978.
24. S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.,* **27**: 113–120, 1979.

25. M. Berouti, R. Schwartz, and J. Makhoul, Enhancement of speech corrupted by acoustic noise. In *Proc. 1979 IEEE ICASSP,* pp. 208–211, April 1979.

26. M. R. Weiss, E. Aschkenasy, and T. W. Parsons, Study and development of the INTEL technique for improving speech intelligibility. Nicolet Scientific Corp., Final Report NSC-FR/4023, Dec. 1974.

27. Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-32**: 1109–1121, 1984.

28. O. Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Proc.,* **2**: 345–349, 1994.

29. R. J. McAulay and M. L. Malpass, Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-28**: 137–145, April 1980.

30. J. H. L. Hansen, A new speech enhancement algorithm employing acoustic endpoint detection and morphological based spectral constraints. In *Proc. 1991 IEEE ICASSP,* pp. 901–904, 1991.

31. T. L. Peterson and S. T. Boll, Acoustic noise suppression in the context of a perceptual model. In *Proc. 1981 IEEE ICASSP,* pp. 1086–1088, April 1981.

32. 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. 1978 IEEE ICASSP,* pp. 602–605, April 1978.

33. R. D. Preuss, A frequency domain noise cancellation preprocessor for narrowband speech communications systems. In *Proc. 1979 IEEE ICASSP,* pp. 212–215, April 1979.

34. C. K. Un and K. Y. Choi, “Improving LPC analysis of noisy speech by autocorrelation subtraction method. In *Proc. 1981 IEEE ICASSP,* pp. 1082–1085, April 1981.

35. G. Whipple, Low residual noise speech enhancement utilizing time-frequency filtering. In *Proc. IEEE ICASSP-94,* vol. I, pp. I5–I8, 1994.

36. P. Lockwood and J. Boudy, Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Commun.,* **11**: 215–228, 1992.

37. L. Arslan, A. McCree, and V. Viswanathan, New methods for adaptive noise suppression. In *Proc. 1995 IEEE ICASSP,* pp. 812–815, 1995.

38. B. George, Single sensor speech enhancement using a soft-decision/variable attenuation algorithm. In *Proc. 1995 IEEE ICASSP,* pp. 816–819, May 1995.

39. D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, Speech enhancement using psychoacoustic criteria. In *Proc. 1993 IEEE ICASSP,* pp. 359–362, April 1993.

40. N. Virag, Speech enhancement based on masking properties of the auditory system. In *Proc. 1995 IEEE ICASSP,* pp. 796–799, May 1995.

41. B. A. Hanson, D. Y. Wong, and B. H. Juang, Speech enhancement with harmonic synthesis. In *Proc. 1983 IEEE ICASSP,* pp. 24.2.1–4, April 1983.

42. D. G. Childers and C. K. Lee, Co-channel speech separation. In *Proc. 1987 IEEE ICASSP,* pp. 181–184, April 1987.

43. J. A. Naylor and S. F. Boll, Techniques for suppression of an interfering talker in co-channel speech. In *Proc. 1987 IEEE ICASSP,* pp. 205–208, April 1987.

44. D. Morgan, E. B. George, L. T. Lee, and S. M. Kay, “Co-channel speaker separation by harmonic enhancement and suppression,” *IEEE Trans. Speech Audio Process.,* **5** (5): 407–424, 1997.

45. J. S. Lim and A. V. Oppenheim, All-pole modeling of degraded speech. *IEEE Trans. Acoust. Speech Signal Process.,* **26**: 197–210, 1978.

46. J. H. L. Hansen and M. A. Clements, Iterative speech enhancement with spectral constraints. In *Proc. 1987 IEEE ICASSP,* pp. 189–192, April 1987.

47. B. R. Musicus and J. S. Lim, Maximum likelihood parameter estimation on noisy data. In *Proc. 1979 IEEE ICASSP,* pp. 224–227, April 1979.

48. J. S. Lim, *Two-Dimensional Signal and Image Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1990.

49. A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *Ann. R. Stat. Soc.,* pp. 1–38, Dec. 1977.

50. M. Feder, A. Oppenheim, and E. Weinstein, Methods for noise cancellation based on the EM algorithm. In *Proc. 1987 IEEE ICASSP,* pp. 201–204, April 1987.

51. M. Feder, A. Oppenheim, and E. Weinstein, Maximum-likelihood noise cancellation in microphones using estimate-maximize algorithm. *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-37**: 1846–1856, 1989.

52. B. R. Musicus, An iterative technique for maximum likelihood parameter estimation on noisy data. S. M. thesis, MIT, Cambridge, Mass., June 1979.

53. K. K. Paliwal and A. Basu, A speech enhancement method based on Kalman filtering. In *Proc. 1987 IEEE ICASSP,* pp. 177–180, April 1987.

54. J. D. Gibson and B. Koo, Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Signal Process.,* **39**: 1732–1742, 1991.

55. S. Nandkumar and J. H. L. Hansen, Dual-channel speech enhancement with auditory spectrum based constraints. In *Proc. 1992 IEEE ICASSP,* pp. 297–300, March 1992.

56. S. Nandkumar and J. H. L. Hansen, Dual-channel iterative speech enhancement with constraints based on an auditory spectrum, *IEEE Trans. Speech Audio Process.,* **3**: 22–34, 1995.

57. J. H. L. Hansen and S. Nandkumar, Robust estimation of speech in noisy backgrounds based on aspects of the auditory process. *J. Acoust. Soc. Am.,* **97** (7): 3833–3849, 1995.

58. D. O'Shaughnessy, Speech enhancement using vector quanitation and a formant distance measure. In *Proc. 1988 IEEE ICASSP,* pp. 549–552, May 1988.

59. J. D. Gibson, T. R. Fisher, and B. Koo, Estimation and vector quantization of noisy speech. In *Proc. 1988 IEEE ICASSP,* pp. 541–544, May 1988.

60. Y. Ephraim, D. Malah, and B. H. Juang, On the application of hidden Markov models for enhancing noisy speech. In *Proc. 1988 IEEE ICASSP,* pp. 533–536, May 1988.

61. Y. Ephraim, D. Malah, and B. H. Juang, Speech enhancement based upon hidden Markov modeling. In *Proc. 1989 IEEE ICASSP,* pp. 353–356, May 1989.

62. J. H. L. Hansen and L. Arslan, Markov model based phoneme class partitioning for improved constrained iterative speech enhancement. *IEEE Trans. Speech Audio Process.,* **3** (1): 98–104, Jan. 1995.

63. B. Widrow, J. R. Grover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, Adaptive noise canceling: Principles and applications. *Proc. IEEE,* **63**: 1692–1716, 1975.

64. J. M. McCool et al., *Adaptive line enhancer,* U.S. Patent No. 4,238,746, December 9, 1980.

65. J. M. McCool et al., *An adaptive detector,* U.S. Patent No. 4,243,935, January 6, 1981.

66. M. M. Sondhi, An adaptive echo canceller. *Bell Syst. Tech. J.,* **46**: 497–511, 1967.

67. J. L. Kelly and R. F. Logan, *Self-adaptive echo canceller,* U.S. Patent No. 3,500,000, March 10, 1970.

68. M. M. Sondhi, *Closed loop adaptive echo canceller using generalized filter networks,* U.S. Patent No. 3,4999,999, March 10, 1970.

69. S. Haykin, *Adaptive Filter Theory,* 2nd. ed., Englewood Cliffs, NJ: Prentice-Hall, 1996.

70. D. G. Messerschmitt, *Adaptive Filters.* Norwell, MA: Kluwer, 1984.

71. B. Widrow and M. E. Hoff, Adaptive switching circuits. In *IRE WESCON Convention Records,* pp. 96–104, 1960.

72. B. Widrow, P. Mantey, L. Griffiths, and B. Goode, Adaptive antenna systems. *Proc. IEEE,* **55**: 2143–2159, 1967.

73. B. Widrow, P. Mantey, L. Griffiths, and B. Goode, Adaptive filters. In R. Kalman and N. DeClaris (eds.), *Aspects of Network and System Theory.* New York: Rinehart and Winston 1971, pp. 563–587.

74. B. Widrow and S. D. Sterns, *Adaptive Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1985.

75. S. F. Boll and D. C. Pulsipher, Suppression of acoustic noise in speech using two microphone adaptive noise cancellation. *IEEE Trans. Acoust. Speech Signal Process.,* **28**: 751–753, 1980.

76. B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, Jr., Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE,* **64**: 1151–1162, 1976.

77. L. J. Griffiths, An adaptive lattice structure for noise-canceling applications. In *Proc. 1978 IEEE ICASSP,* pp. 87–90, 1978.

78. W. A. Harrison, J. S. Lim, and E. Singer, A new application of adaptive noise cancellation. *IEEE Trans. Acoust., Speech, Signal Process.,* **34**: 21–27, 1986.

79. G. A. Powell, P. Darlington, and P. D. Wheeler, Practical adaptive noise reduction in the aircraft cockpit environment. In *Proc. 1987 IEEE ICASSP,* pp. 173–176, April 1987.

80. J. J. Rodriguez, J. S. Lim, and E. Singer, Adaptive noise reduction in aircraft communication systems. In *Proc. 1987 IEEE ICASSP,* pp.169–172, April 1987.

81. M. R. Sambur, Adaptive noise canceling for speech signals. *IEEE Trans. Acoust. Speech Signal Process.,* **26**: 419–423, 1978.

82. M. R. Sambur, LMS adaptive filtering for enhancing the quality of noisy speech. In *Proc. 1978 IEEE ICASSP,* pp. 610–613, April 1978.

83. L. W. Varner, T. A. Miller, and T. E. Eger, A simple adaptive filtering technique for speech enhancement. In *Proc. 1983 IEEE ICASSP,* pp. 24.3.1–4, April 1983.

84. J. W. Kim and C. K. Un, Enhancement of noisy speech by forward/backward adaptive digital fitlering. In *Proc. 1986 IEEE ICASSP,* pp. 89–92, April 1986.

85. J. S. Lim and A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech. *Proc. IEEE,* **67**: 1586–1604, 1979.

86. D. Malah and R. V. Cox, Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Trans. Acoust. Speech Signal Process.,* **27**: 121–133, 1979.

87. D. Malah and R. V. Cox, A generalized comb filtering technique for speech enhancement. In *Proc. 1982 IEEE ICASSP,* pp. 160–163, 1982.

88. R. V. Cox and D. Malah, A technique for perceptually reducing periodically structured noise in speech. In *Proc. 1981 IEEE ICASSP,* pp. 1089–1092, April 1981.

89. Y. M. Perlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, Evaluation of a speech enhancement system. In *Proc. 1977 IEEE ICASSP,* pp. 212–215, May 1977.

90. J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for evaluating speech degraded by white noise addition. *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-26**: 354–358, 1978.

91. Y. M. Cheng and D. O'Shaughnessy, Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Signal Process.,* **39**: 1943–1954, 1991.

92. D. E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, Speech enhancement based on audible noise suppression. *IEEE Trans. Speech Audio Process.,* **5**: 497–514, 1997.

93. J. H. L. Hansen, Analysis and compensation of stressed and noisy speech with application to robust automatic recognition. Ph.D. thesis, Georgia Institute of Technology, July 1988.

94. S. F. Boll, Adaptive noise canceling in speech using the short-time transform. In *Proc. 1980 IEEE ICASSP,* pp. 692–695, April 1980.

95. A. V. Oppenheim et al., Single-sensor active noise cancellation, *IEEE Trans. Speech Audio Process,* **2** (2): 285–290, April 1994.

96. K. C. Zangi, A new two-sensor active noise cancellation algorithm, In *Proc. 1993 IEEE ICASSP,* pp. II-351–354, April 1993.

97. S. J. Elliott and P. A. Nelson, Active noise control, *IEEE Signal Proc. Mag.,* **10** (4): 12–35, Oct. 1993.

98. H. Sheikhzadeh, R. L. Brennan, and H. Sameti, Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications, *Proc. 1995 IEEE ICASSP,* 808–811, May 1995.

JOHN H. L. HANSEN
Duke University

SPEECH, LANGUAGE IDENTIFICATION. See AUTOMATIC LANGUAGE IDENTIFICATION.