

Speech Enhancement for Crosstalk Interference

Levent M. Arslan, *Member, IEEE*, and John H. L. Hansen, *Senior Member, IEEE*

Abstract—Speech signals are often degraded by additive interference over single-channel communication systems. For stationary and well defined noise sources, effective solutions exist. However, it is often difficult to formulate a model for nonstationary and speechlike noise sources such as crosstalk or multispeaker babble, which exist in real scenarios. In this paper, we propose a solution to this problem under the assumption that we have access to the clean speech signal prior to transmission. A novel method for tracking transmission noise characteristics is described. Based on this noise estimate, a new speech enhancement technique is proposed. The enhancement method is evaluated for multispeaker babble noise, and shown to substantially improve both the quality and intelligibility of the processed speech signal.

I. INTRODUCTION

IN SINGLE-CHANNEL voice communication systems, it is often difficult to characterize background interfering noise. The channel distortion normally possesses nonstationary statistics, and can contain correlated interference (e.g., another speaker's voice). However, most models developed for single-channel speech enhancement systems assume that background noise is stationary and/or uncorrelated [1], [2], [4], [7]. Although the fundamental principles behind these enhancement methods are well defined, in practice, the limitations set by their assumptions play a major role in their performance across actual distortions. The reason for this is that they rely on a good estimate of the noise characteristics, which can have serious consequences when the assumptions are violated. Another limitation of traditional methods is that they rely on a short-time stationarity assumption, which is not valid for some speech classes such as stop consonants. As a result, these enhancement algorithms can introduce artifacts which reduce overall speech intelligibility.

Methods have been proposed that seek to preprocess clean speech prior to transmission across a channel in an effort to increase intelligibility [8], [9]. Unfortunately, these methods generally compromise overall speech quality as a result of their processing. In this paper, we propose a time-division multiplexing-based scheme to track the channel noise characteristics without imposing any constraints on the noise type. The proposed method is very simple; however, it requires access to the clean speech signal prior to transmission as in [8] and [9]. The method is based on padding the signal with zeros

Manuscript received August 4, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. V. Viswanathan.

L. M. Arslan was with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA. He is now with Entropic Research, Washington, DC, USA (e-mail: arslan@epiwr12.entropic.com).

J. L. Hansen is with the Robust Speech Processing Laboratory, Department of Electrical Engineering, Duke University, Durham, NC 27708-0291 USA (e-mail: jlh@ee.duke.edu).

Publisher Item Identifier S 1070-9908(97)02521-2.

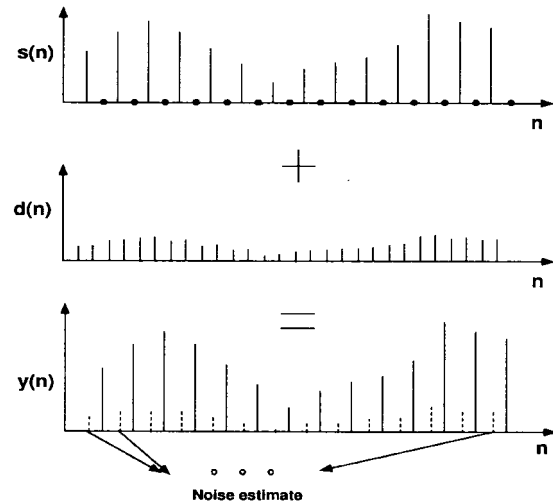


Fig. 1. Zero-padding procedure for accurate noise estimation, where $s(n)$ is the transmitted zero-padded clean speech signal, $d(n)$ is the interference signal that is added in the channel, and $y(n)$ is the output noisy signal at the receiver.

at the transmitter prior to transmission across the channel, and estimating the noise characteristics from the original zero samples that are now degraded when collected at the receiver. Since most noise signals have correlation between successive samples (especially speechlike interference), the noise samples that are added to a signal sample will be very similar to those noise samples that are added to the closest zero sample. Therefore, even if the degraded zero samples from neighboring signal samples are subtracted at the output, the resulting speech will possess both higher quality and intelligibility.

The outline of this paper is as follows: In Section II, the zero-padding procedure for channel noise estimation is presented. In Section III, the evaluations including multispeaker babble noise interference are presented. Finally, in Section IV, the conclusions are presented.

II. ZERO-PADDING PROCEDURE

The procedure for obtaining the noise estimate is shown in Fig. 1, where the top plot shows the transmitted signal $s(n)$, which is padded with zeros at every other sample. The second plot corresponds to the interference signal $d(n)$, which is assumed to be an additive noise distortion due to the channel. The resulting signal at the receiver $y(n)$ is shown at the bottom plot. In this procedure, the noise is estimated from the original zero samples which are marked with dashed lines. One approach for enhancing the output signal is to simply subtract the noise estimate (marked with dashed lines) from the noisy speech signal (marked with solid lines in the bottom plot). This method will be referred to as *sample subtraction*.

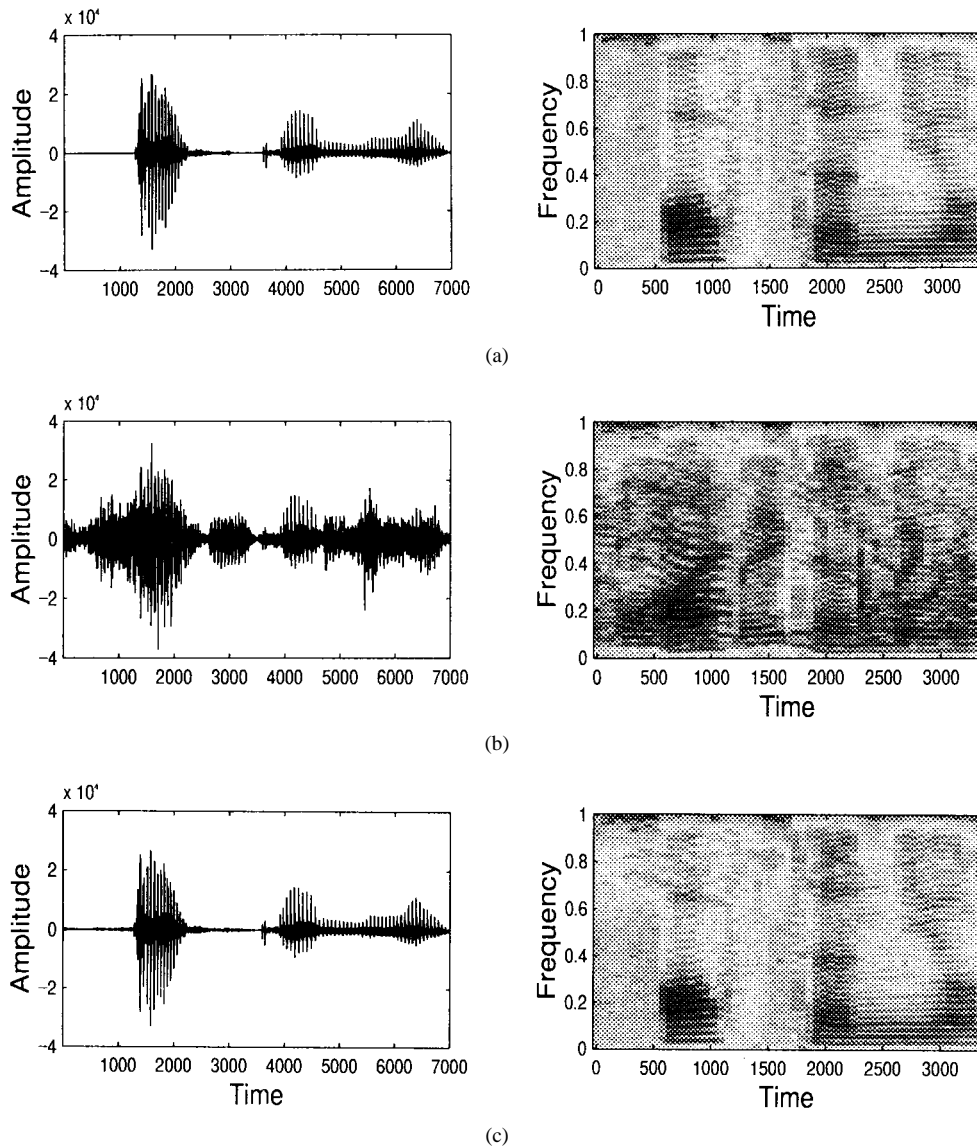


Fig. 2. Time waveforms and spectrograms for the utterance “Often you’ll” from the TIMIT sentence “Often you’ll get back more than you put in.” (a) Original utterance. (b) Degraded with -10 dB multispeaker babble noise. (c) Enhanced using the zero-padding procedure.

To improve upon the enhancement procedure, interpolation techniques can be applied in order to obtain a better estimate of the noise signal that interferes with nonzero samples. It should be noted that the resampling process at the receiver should be synchronized to the transmitter sampling. Small errors in the synchronization process may produce a mix of signal in addition to noise, and this may reduce the performance of the proposed algorithm. Therefore, the synchronization process should be precise.

One disadvantage of the zero-padding sample subtraction procedure is that it requires twice the data rate, or two times the size of the original channel bandwidth for transmission. In order to reduce the bandwidth requirement, zero padding can be based on the degree of correlation between successive noise samples, so that zeros may be padded every second sample, third sample, etc. This will reduce the bandwidth requirement from $2/1$ to $3/2$, $4/3$ times, etc., respectively. However, reducing the bandwidth will result in a less accurate estimate of the noise characteristics. Based on the particu-

lar voice communication application, and available channel bandwidth, an appropriate value can be estimated experimentally.

Another issue is the nonideal filter characteristics of the bandlimited channel. Under ideal conditions, the channel can be modeled as a filter with perfect passband/stopband characteristics, and therefore each sampled pulse of speech spaced $1/f_s$ apart will produce a sinc function $[\sin(x)/x]$ type response, but will still maintain a null at the intermediate point in time between samples. Since these null sample locations correspond to noise estimate samples in our formulation, the ideal channel filter characteristics will not result in distortion of the noise estimate. However, in practice, the channel filter characteristics may not be ideal. This will produce a smearing of the speech pulses that would result in leakage into the zero-valued samples reserved for the noise estimate. This problem can be resolved to some extent by employing an adaptive filter to remove the smeared component of the speech signal from the noise signal. A number of techniques for echo cancellation

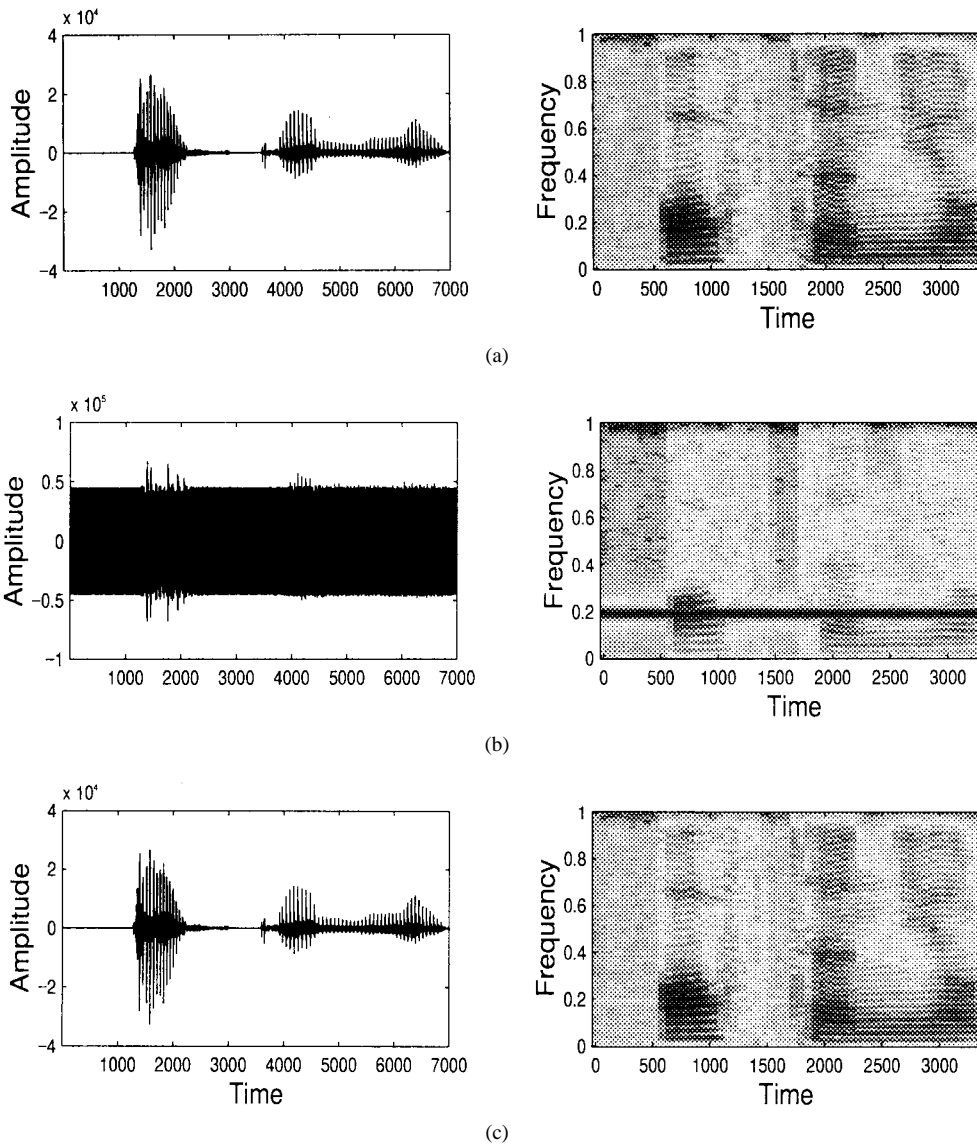


Fig. 3. Time waveforms and spectrograms for the utterance "Often you'll" from the TIMIT sentence "Often you'll get back more than you put in." (a) Original utterance. (b) Degraded with -32 dB sinusoidal interference. (c) Enhanced using the zero-padding procedure.

found in the literature [5], [6] could be employed to address this issue.

It is important to note that the sample subtraction method will be more effective if the successive noise samples are correlated. However, if the successive noise samples are uncorrelated, then speech enhancement could be performed in the frequency domain using one of the traditional approaches, such as spectral subtraction or Wiener filtering, on a frame-by-frame basis. Since both of these methods require a good noise estimate, the proposed noise estimation procedure will increase frequency domain speech enhancement performance as well. One of the most important advantages of the proposed method is that it does not require any stationarity assumption, which is a major problem for existing speech enhancement techniques. The reason for this is that the noise estimate is updated automatically for every other input sample.

The degree of correlation between successive noise samples plays a major role in deciding between sample subtraction or traditional speech enhancement methods for receiver-end

speech enhancement. As mentioned above, for either case, zero-padding-based noise estimation will improve the performance substantially. However, in order to achieve the highest level of performance, a decision mechanism between the two processing methods can be embedded in the speech enhancement structure at the receiver. The criterion for switching between the methods would depend on the degree of correlation between successive noise samples. In order to formulate a mathematical expression for the degree of correlation, we define the sequences X and Y as

$$\begin{aligned} X &= d_n d_{n+1} d_{n+2} \cdots d_{n+N} \\ Y &= d_{n-1} d_n d_{n+1} \cdots d_{n+N-1} \end{aligned} \quad (1)$$

where N is a predefined frame length. Next, the correlation coefficient between X and Y is obtained using the expression

$$\rho = \frac{K_{XY}}{\sigma_X \sigma_Y} \quad (2)$$

where K_{XY} is the covariance, which is defined as follows:

$$K_{XY} = E[(X - m_X)(Y - m_Y)] \quad (3)$$

where m_X and m_Y are the means of X and Y , respectively. The correlation coefficient ρ can be updated every other input sample in order to direct the enhancement decision mechanism as needed.

III. EVALUATIONS

In these evaluations, speech consisted of continuous sentences from the TIMIT speech database, downsampled to an 8 kHz sample rate. For the first evaluation, the proposed speech enhancement method is applied to the problem of enhancing multispeaker babble noise interference [3]. Fig. 2(a) shows the time waveform and corresponding spectrogram for the utterance ‘‘Often you’ll’’ which is part of the TIMIT sentence ‘‘Often you’ll get back more than you put in’’ spoken by a male speaker. Fig. 2(b) corresponds to the degraded waveform and its spectrogram with -10 dB SNR of multiple speaker babble noise. At this level of noise, the original speech signal is not distinguishable. Competing speaker formant tracks are also clearly visible in the adjoining speech spectrogram. However, after applying the proposed method of enhancement, the original clean signal is recovered with virtually no perceived residual noise. A portion of the recovered signal is shown in Fig. 2(c). The mean square error between the degraded signal and the original signal improved from 2653 to 137 after applying the sample subtraction enhancement procedure.

Next, a degrading sinusoidal interference was considered. As previously mentioned, the effectiveness of the algorithm is more pronounced when the noise interference is more highly correlated, which is the case for the sinusoidal interference. Fig. 3(a) shows the original utterance ‘‘Often you’ll.’’ Fig. 3(b) corresponds to the degraded waveform and its spectrogram with -30 dB sinusoidal interference at 700 Hz. At this level of noise, listener evaluation indicates that only the single tone is heard, and virtually no speech signal can be perceived. However, after applying the proposed enhancement method, the original signal is completely recovered, as can be seen in Fig. 3(c). Here the mean square error drops from 47 750 to 14 after applying the sample subtraction enhancement procedure.

IV. CONCLUSIONS

In this letter, a new method for estimating degrading noise characteristics was proposed, and integrated into a speech enhancement scheme. Our proposed method assumed access to the clean speech signal prior to transmission. The method is based on simply padding the signal with zeros at every other sample in order to characterize the background noise in the communications system at the receiver. Using the proposed method, it has been shown that the original speech can be easily reconstructed in the presence of such noise sources as multispeaker babble noise or sinusoidal interference. The usage of the method is illustrated here for nonstationary and correlated noise types, since this noise type normally causes traditional speech enhancement algorithms to fail. In closing, it should be mentioned that the method is flexible enough to accommodate many typical noise sources, and quite appropriate for real-time implementation.

REFERENCES

- [1] L. M. Arslan, A. McCree, and V. Viswanathan, ‘‘New methods for adaptive noise suppression,’’ in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Detroit, MI, May 1995, vol. 1, pp. 812–815.
- [2] S. F. Boll, ‘‘Suppression of acoustic noise in speech using spectral subtraction,’’ *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, April 1979.
- [3] J. H. L. Hansen and L. M. Arslan, ‘‘Robust feature estimation and objective quality assessment for noisy speech recognition using credit card corpus,’’ *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 169–184, May 1995.
- [4] J. H. L. Hansen and M. A. Clements, ‘‘Constrained iterative speech enhancement with application to speech recognition,’’ *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 795–805.
- [5] S. Haykin, *Adaptive Filter Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.
- [6] J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
- [7] J. S. Lim and A. V. Oppenheim, ‘‘All-pole modeling of degraded speech,’’ *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, pp. 197–210, 1978.
- [8] R. J. Niederjohn and J. H. Grotelueschen, ‘‘The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression,’’ *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 277–282, Aug. 1976.
- [9] I. B. Thomas and R. J. Niederjohn, ‘‘The intelligibility of filtered-clipped speech in noise,’’ *J. Audio Eng. Soc.*, vol. 18, pp. 299–303, June 1970.