Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition

JAE S. LIM, STUDENT MEMBER, IEEE, ALAN V. OPPENHEIM, FELLOW, IEEE, AND LOUIS D. BRAIDA, MEMBER, IEEE

Abstract—**An intelligibility test was performed to evaluate an adaptive comb filtering method proposed by Frazier [2] for enhancement of degraded speech due to additive white noise. Results indicate that independent of S/N ratio the adaptive comb filtering scheme does not increase speech intelligibility.**

I. Introduction

THERE are a variety of practical situations in which enhancement of speech degraded by additive interfering signals is desirable. One approach which has been considered in detail [1]-[5] is based on the observation that waveforms of voiced sounds are periodic with a period that corresponds to the fundamental frequency. Making use of this knowledge, a comb filtering operation that passes only the harmonics of speech was applied by Shields [1] to enhance degraded speech. Since interfering signals will, in general, have energy in the frequency regions between the speech harmonics, this operation in principle can reduce noise while preserving speech signals to the extent that information of the fundamental frequency is available and the periodicity of speech is strictly preserved. Frazier [2], [3] observed that even with accurate fundamental frequency information, Shields' comb filtering method distorts speech signals significantly due to the time varying nature of speech sounds. To reduce some of this distortion, Frazier suggested an adaptive comb filter which adjusts itself to variations in fundamental frequency. Using Frazier's system, Perlmutter *et al.* [4], [5] processed some speech material and performed intelligibility tests with interference consisting of the speech of a single competing talker. Her results indicate that even with accurate fundamental frequency information, adaptive comb filtering decreases intelligibility for S/N ratios in the range of -3 to 9 dB.

The adaptive comb filtering method had not yet previously been applied to the problem of broad-band random noise interference. This paper reports the results of some intelligibility tests conducted for speech material corrupted by broadband random noise using a modified version of Frazier's adaptive comb filter.

II. Principles of an Adaptive Comb Filter

The basic goal of adaptive comb filtering is noise reduction without speech distortion. Since detailed explanations and problems associated with an adpative comb filter can be found in Frazier [2], only a brief description is given in this section.

The operation of an adaptive comb filter can be explained by considering its unit sample response over one pitch period:

$$h(n) = \sum_{k=-L}^{L} a_k \cdot \delta(n - N_k). \tag{1}$$

Here, $h(n)$ is the unit sample response, $\delta(n)$ is a unit sample function, the length of the filter is $2L + 1$ pitch periods, a_k is the filter coefficient that satisfies $\Sigma_{k=-L}^{L} a_k = 1$, and N_k is given by the following equations:

$$N_k = \begin{cases} -\sum_{l=k}^{-1} T_l, & \text{for } k < 0 \\ 0, & \text{for } k = 0 \\ \sum_{l=0}^{k-1} T_l, & \text{for } k > 0, \end{cases} \tag{2}$$

where T_k corresponds to the particular pitch period which contains the point of speech waveform that is multiplied by the filter coefficient a_k. Except for a few instances that will be discussed in the next section, the filter coefficients are unchanged and only N_k is updated once every pitch period based on pitch information of the speech waveform being processed. The specific filter coefficients that were used in processing test materials will be discussed in the next section. An example in which the filter is 5 pitch periods long is shown in Fig. 1.

To the extent that the speech waveform is periodic over the $2L + 1$ pitch periods that the filter is applied, the speech samples will add constructively, but the noise samples will tend to sum toward zero.

III. Algorithm Used for Processing Test Materials

The algorithm used in processing speech for the intelligibility test is illustrated schematically in Fig. 2. In considering the algorithm in Fig. 2, the following points should be noted.

First, the fundamental frequency information including the

Manuscript received April 5, 1977; revised July 27, 1977. This work was supported by the Advanced Research Projects Agency monitored by ONR under Contract N00014-75-C-0951-NR 049-308.

The authors are with Department of Electrical Engineering and Computer Science, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

0096-3518/78/0800-0354$00.75 © 1978 IEEE

voiced/unvoiced decision[1] was obtained from the glottal waveform which was measured at the time when speech was recorded. This information was hand corrected with reference to the acoustical signal to obtain accurate estimates of glottal waveform period.

Second, digitally generated white noise was used as a degrading source. Each noise sample was obtained from a Gaussian density function (restricted to lie within ±3.5 standard deviations) and made to be statistically independent of any other sample. The noise amplitude was adjusted to achieve specified S/N (speech to noise)[2] ratios for the intelligibility test.

Third, for sections of speech that correspond to unvoiced sounds or silence, simple attenuation was applied. This step is necessary because applying an adaptive comb filter to voiced sounds reduces the noise energy that is added to voiced sounds. Not applying attenuation to speech sections corresponding to unvoiced sounds or silence has the effect of unnatural emphasis for unvoiced sounds or silence relative to voiced sounds.

Fourth, the filter coefficients a_k which determine the unit sample response of the adaptive comb filter $h(n)$ were chosen from qualitative results that Frazier [2] obtained from informal listening tests. Frazier considered four different kinds of unit sample response shapes corresponding to rectangular, Hamming, Hanning, and Blackman windows. He reported that the adaptive comb filter with a rectangular shape produced lowest intelligibility while there was little difference in intelligibility for speech sounds produced by the other three types of filters. The filter that has been used in this paper has a Hamming window shape which is obtained from the following equation[3]:

$$a_k = \frac{0.54 + 0.46 \cos (2\pi k/2L + 1)}{\sum_{k=-L}^{L} 0.54 + 0.46 \cos (2\pi k/2L + 1)}, \quad \text{for } -L \leqslant k \leqslant L. \quad (3)$$

The sentences that were processed for the intelligibility test are for the cases when $L = 1, 3,$ and 6.[4]

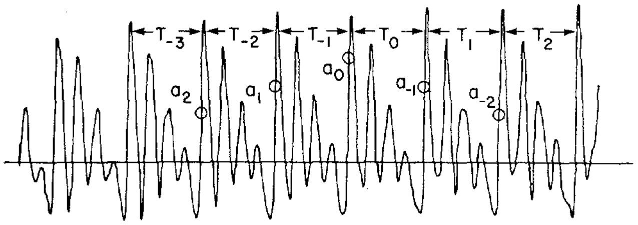

Fig. 1. Example of an adpative comb filter.

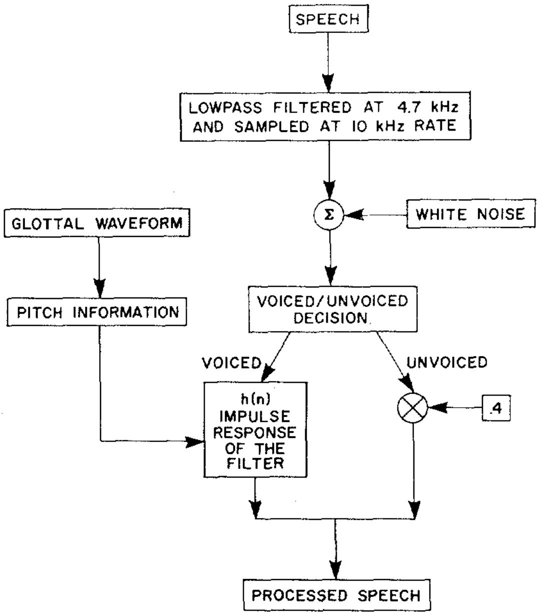

Fig. 2. Adaptive comb filtering algorithm used for processing test material.

Fifth, in processing voiced sounds, there are two situations under which some portion of the filter is turned off. The first situation, referred to as the "overload problem,"[5] is illustrated by the example of an adaptive comb filter which was shown in Fig. 1. In the figure, when T_0 is longer than any of $T_i (i \neq 0)$, there exist times when more than one pitch period is covered by some filter coefficient a_k while a_0 is applied to the full length of T_0. For such cases, the value of a_k was made to be zero for the portion that exceeds one local pitch period. The second situation arises when a transition between voicing and unvoicing occurs. When voiced sounds near the transition are processed, the adaptive comb filter extends over unvoiced sounds due to the filter length. In such cases, the filter coefficients that extend over unvoiced sounds are set to zero. This particular method of treating voiced sounds near transition differs from Frazier's algorithm in which the filter is unchanged.[6] This modification, which will be discussed in a later section, noticeably reduced speech distortion near the points of transition between voiced and unvoiced sounds. In both cases when

[1] Since the processing algorithm is the same for both unvoiced sound and silence, silence was treated as unvoiced.

[2] The S/N ratio that determines the noise power added to speech is related to speech and noise as follows: for each test sentence

$$S/N \text{ in dB} = 10 \log \frac{\sum_n s^2(n)}{\sum_n w^2(n)}$$

where $s(n)$ is speech waveform, $w(n)$ is noise waveform, and summation is over the length of the test sentence.

[3] The denominator of (3) for a_k guarantees the condition that

$$\sum_{k=-L}^{L} a_k = 1.$$

[4] The length of the adaptive comb filter is about $2L + 1$ pitch periods long. Hence, $L = 1$, 3, and 6 correspond to the filter lengths of 3, 7, and 13 pitch periods.

[5] A more detailed description of "overload problem" can be found in Frazier [2].

[6] In Frazier's algorithm, the pitch periods assumed for filter coefficients that extend over unvoiced sounds are the pitch period of a voiced sound nearest the point of transition between voicing and unvoicing. This method was used by Perlmutter *et al.* [4], [5].

TABLE I
INTELLIGIBILITY SCORES IN PERCENT OF CORRECT RESPONSES

Filter Length S/N in dB	no processing	3 pitch periods	7 pitch periods	13 pitch periods
∞	98.9	98.4	94.3	81.3
5	64.2	61.7	45.6	27.5
0	51.4	52.7	32.3	19.4
-5	25.6	27.1	13.1	11.3

some portion of the filter is turned off, the remaining coefficients are scaled up linearly such that the sum of the nonzero coefficients remains unity.

IV. Test Material and Procedures

In general, a fair evaluation of a speech enhancement system should be based on many factors such as intelligibility, fatigue, cost of implementation, etc. In this paper, we take a very limited view and consider primarily the effect of the system on intelligibility of wide-band speech.

The test material and procedures are essentially the same as those used in the intelligibility test by Perlmutter *et al.* [4], [5].

The test material consists of eighty nonsense sentences which are in the format of "The *adjective noun verb* (past tense) the *noun*." One such example is: "The *strong ball built* the *rock*." The procedure used for sentence construction is similar to one proposed by Nye and Gaitenby [6]. The sentences were constructed from a random selection of monosyllabic words from a list of 63 adjectives, 63 verbs in the past tense, and 126 nouns. The list of words was drawn from the first 2000 words in the Thorndike and Lorge [7] count. Twenty such sentences constructed in this manner were each spoken by two female and two male young adults.

The S/N ratios used in the test are -5 dB, 0 dB, 5 dB, and ∞.[7] These values were chosen on the basis of preliminary informal tests to obtain scores that range between 20 and 100 percent.

A session consisted of a practice test and a main test. A main test consisted of presentation of a total of eighty processed or unprocessed sentences of one S/N ratio. Each listener was allowed to join a maximum of two sessions and a total of twenty five listeners participated in the tests.

V. Data Analysis and Results

Responses were graded for the number of words correctly recorded. The rule in grading is that a recorded word is considered to be correct only when the pronounced sound of the recorded word is identical to that of the correct word. All other cases were considered to be wrong.

The data obtained from grading were classified as a function of the experimental parameters. Data in a given classification were then combined together to obtain an intelligibility score which is the percentage of correct answers. Since the experimental parameters consisted of four different S/N ratios (-5 dB, 0 dB, 5 dB, and ∞) and four different filter lengths (no processing, 3 pitch periods, 7 pitch periods, 13 pitch periods), there were 16 possible combinations. The intelligibility score for each of these combinations is given in Table I and plotted in Fig. 3.

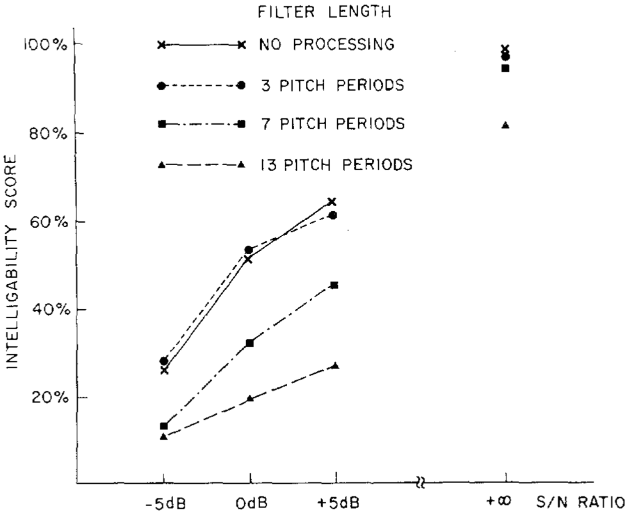

Fig. 3. Plot of intelligibility scores in percentage of correct responses. Each point corresponds to roughly 500 responses pooled over 7 listeners.

VI. Discussion

As is expected, results from the intelligibility test show that intelligibility score for a given filter length decreases as S/N ratio decreases. Results also show that intelligibility score for a given S/N ratio generally decreases as the filter length increases.

Results in this study indicate that even with perfect information of fundamental frequency the adaptive comb filter does not achieve any significant increase of intelligibility at any S/N ratio of interest when the degrading source is additive white noise. In fact, a substantial decrease of intelligibility is observed when the filter lengths are 7 and 13 pitch periods. Only when the filter is 3 pitch periods long, intelligibility score does not show any noticeable decrease. If pitch period information has to be derived from degraded speech as would be the case in practical situations, intelligibility score will be

[7] S/N ratio = ∞ means no noise was added to speech.

even worse. These results lead to the conclusion that the kind of adaptive comb filter that has been considered in this paper is not effective in increasing speech intelligibility when the degrading source is additive white noise.

Qualitatively speaking, these results are consistent with those of Perlmutter [4]. There are some detailed differences, however, in the results between the two studies. When the S/N ratio $= \infty$, results of the two studies are directly comparable since the same test materials were used in both studies. Results in this study indicate that there is little difference in intelligibility score between speech with no processing and speech processed by an adaptive comb filter with length of 3 pitch periods, while results by Perlmutter indicate that the score decreases by about 4 percent. For other filter lengths (7 and 13 pitch periods) at S/N ratio $= \infty$, similar observations can be made, which implies that speech distortion caused by the algorithm used in this study is less than speech distortion by the algorithm used by Frazier and Perlmutter.[8]

It may be possible to make some improvements in the algorithm of the filter that has been considered in this paper. For example, the above algorithm is based on the assumption that waveforms of voiced sounds do not vary rapidly from one pitch period to the next. A more realistic assumption is that the mechanical attributes of a human vocal tract, such as the cross sectional area function rather than the waveforms themselves, do not vary significantly from one pitch period to the next pitch period. If we can obtain such attributes from speech waveforms, then we can apply the adaptive comb filtering concept to such attributes rather than to the speech waveforms. This aspect of improving the algorithm is under investigation.

If increasing speech intelligibility is not the only objective of processing, the adaptive comb filter considered in this paper may still have some area of application due to its capability of noise reduction. For voiced sounds, the approximate S/N ratio increase due to the adaptive comb filter can be obtained in the following manner. The output of an adaptive comb filter considered in this paper can be represented by

$$y(n) = \sum_{k=-L}^{L} a_k \cdot x(n - N_k) \tag{4}$$

where $x(n)$ is degraded speech, $y(n)$ is processed speech, and N_k is the point of filter coefficient a_k. Since $x(n) = s(n) + w(n)$ where $s(n)$ is speech and $w(n)$ is degrading source, we have

$$y(n) = \sum_{k=-L}^{L} a_k \cdot s(n - N_k) + \sum_{k=-L}^{L} a_k \cdot w(n - N_k). \tag{5}$$

Assuming $s(n) = \Sigma_{k=-L}^{L} a_k \cdot s(n - N_k)$, which is the basis for the adaptive comb filter,

$$y(n) = s(n) + \sum_{k=-L}^{L} a_k \cdot w(n - N_k). \tag{6}$$

From (6)

S/N ratio in dB for $y(n)$

$$= 10 \log \frac{\sum_n s^2(n)}{E\left[\sum_n \left(\sum_{k=-L}^{L} a_k \cdot w(n - N_k)\right)^2\right]}$$

$$= 10 \log \frac{\sum_n s^2(n)}{\sum_n \sum_{k=-L}^{L} a_k^2 \cdot N_0} \tag{7}$$

where

$N_0 = E[w^2(n)]$.

Since $x(n) = s(n) + w(n)$,

$$S/N \text{ ratio in dB for } x(n) = 10 \log \frac{\sum_n s^2(n)}{\sum_n N_0}. \tag{8}$$

From (7) and (8),

S/N ratio increase in dB due to processing

$= S/N$ ratio in dB for $y(n) - S/N$ ratio in dB for $x(n)$

$$= -10 \log \sum_{k=-L}^{L} a_k^2. \tag{9}$$

Applying (9) to the cases of three filter lengths in this study, we obtain the results shown in Table II. The above results show that the approximate S/N ratio increase due to the adaptive comb filter for the filter lengths of 3, 7, and 13 pitch periods is 3.5 dB, 7 dB, and 10 dB, respectively. This is consistent with the empirical observation that the processed speech "sounds" less noisy. Thus, the adaptive comb filter may be useful for the objectives where noise reduction without significant decrease in speech intelligibility is important.

Acknowledgment

We appreciate the help of Y. Perlmutter for supplying files of nonsense sentences and their fundamental frequency information which were used in the processing. Our appreciation also goes to Dr. B. Gold and Mr. N. Durlach for useful suggestions for the intelligibility test.

References

[1] V. C. Shields, Jr., "Separation of added speech signals by digital comb filtering," S. M. thesis, M.I.T., Cambridge, 1970.

[2] R. H. Frazier, "An adaptive filtering approach toward speech enhancement," S. M. thesis, M.I.T., Cambridge, 1975.

[3] R. H. Frazier, S. Samsam, L. D. Braida, and A. V. Oppenheim,

[8] The difference in the two algorithms is in the treatment of voiced sounds near points of transition between voicing and unvoicing. This was discussed in Section III.

TABLE II
APPROXIMATE S/N RATIO INCREASE IN dB DUE TO PROCESSING BY THE ADAPTIVE COMB FILTERS CONSIDERED IN THIS PAPER

Filter Length in Pitch Periods	a_{-6}	a_{-5}	a_{-4}	a_{-3}	a_{-2}	a_{-1}	a_0	a_1	a_2	a_3	a_4	a_5	a_6	S/N Ratio Increase in dB
3	.0	.0	.0	.0	.0	.191	.617	.191	.0	.0	.0	.0	.0	3.427 dB
7	.0	.0	.0	.033	.116	.219	.265	.219	.116	.033	.0	.0	.0	7.107 dB
13	.013	.028	.054	.085	.114	.135	.142	.135	.114	.085	.054	.028	.013	9.795 dB

"Enhancement of speech by adaptive filtering," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Philadelphia, PA, Apr. 12-14, 1976, pp. 251-253.

[4] Y. M. Perlmutter, "Evaluation of a speech enhancement system," S. M. thesis, M.I.T., Cambridge, 1976.

[5] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, and A. V. Oppenheim, "Evaluation of a speech enhancement system," presented at the 1977 IEEE Int. Conf. on Acoust., Speech, and Signal Processing, Hartford, CT, May 9-11, 1977.

[6] P. W. Nye and J. H. Gaitenby, "The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences," *Status Rep. Speech Res.*, Haskins Lab., pp. 169-190, 1974.

[7] E. L. Thorndike and I. Lorge, *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, 1968.

Real-Time Harmonic Pitch Detector

STEPHANIE SENEFF

***Abstract*—A real-time harmonic pitch detection algorithm has been developed on the Lincoln Digital Voice Terminal (LDVT). The algorithm was designed to be fast and to perform well when the input speech is degraded (i.e., telephone quality) or corrupted with acoustically coupled noise. The algorithm determines the fundamental frequency from the spacing between harmonics in a selected portion of the spectrum. The algorithm was incorporated into a real-time linear prediction vocoder and compared favorably in informal listening tests with the Gold-Rabiner time-domain detector under a variety of adverse conditions.**

INTRODUCTION

THE speech waveform can be modeled as the response of the vocal tract filter to a source which is a periodic sequence of pulses during voiced segments or a random noise during unvoiced segments. The periodic pulses occur as a consequence of the opening and closing of the glottis, and the frequency of the periodicity is often referred to as the pitch.[1] The noise source is a consequence of a narrow constriction at some point in the vocal tract. The model is a simplification, for certain sounds, such as /v/, are driven by both a periodic and a noise source simultaneously. However, the model has proved to be sufficiently accurate that present-day vocoders are constructed based on the simple concept of either a periodic or noise source, but not both, at a given instance of time.

Once the model has been accepted, the difficult task is to determine the periodicity of the source when the speech is voiced, and to determine that there is no periodicity when it is unvoiced. It is a task to which considerable attention has been devoted in the past, and as a consequence there are many published papers available on the subject of pitch extraction [2]-[8], [10], [11], [13].

Pitch detection is generally good when the input signal is intact and noise-free. However, distortions, filters, and noise tend to obscure the pitch information and cause most pitch detectors to break down, sometimes severely. Since in the real world the signal is often corrupted, it was felt that an algorithm designed to be robust against degradations would be a significant new contribution.

We were particularly interested in coping with degradations of the type caused by passage of the speech through the public telephone system prior to pitch detection. One approach to evaluation of a pitch detector on telephone quality speech would be to dial up a few lines at random and analyze performance on the resulting degraded speech. However, given the immense variability of the quality of telephone connections, one has no idea of 1) how representative the selected lines are or 2) what their characteristics are in terms of noise, filtering, and distortions. A more systematic approach would be to test the algorithm's performance on speech first processed through a telephone channel simulation. A disadvantage of such a

Manuscript received June 22, 1977; revised November 16, 1977 and March 22, 1978. This work was sponsored by the Advanced Research Projects Agency of the Department of Defense.

The author is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173

[1] Pitch is more strictly defined as the perceived rather than the generated frequency; the latter has been given various names, such as, laryngeal frequency, voice fundamental frequency, etc., but all are rather awkward compared to "pitch."