A TECHNIQUE FOR PERCEPTUALLY REDUCING PERIODICALLY STRUCTURED NOISE IN SPEECH

R. V. Cox
*D. Malah**

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

Periodically structured noise is noise which occurs randomly but with a fixed or slowly varying period. The noise periodicity is usually due to some underlying process, such as block processing of the speech where discontinuities between successive blocks result. This type of noise permeates the entire speech spectrum and is not removable by standard filtering techniques. The recently developed time domain harmonic scaling (TDHS) algorithm has been found to be the basis for an effective enhancement technique. In this paper we discuss the underlying theory of this technique and establish a class of windows for its implementation. As an example the frame rate noise of adaptive transform coding was perceptually reduced using this technique. Results from a subjective testing experiment using ATC coded speech with bit rates of 7.2 to 16 Kb/s indicated an improvement in quality equivalent to an increase in code rate of 2.4 to 3 Kb/s for speech originally coded at 7.2 to 12 Kb/s.

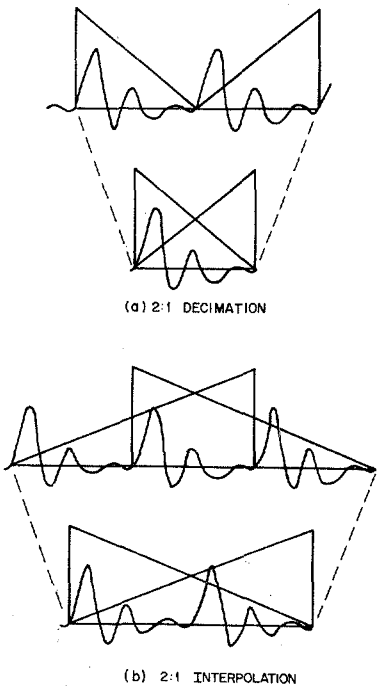
Introduction

In some applications speech is corrupted by an underlying process with a periodic nature. The resulting noise will have a periodic structure. An example of a noise which falls into this category is artifacts caused by block processing of the speech. Because the noise is periodic its spectrum will also be periodic. This means the noise spectrum will permeate the entire speech spectrum. This makes it very difficult to entirely remove the noise. If the noise is low level in comparison to the speech, properties of auditory masking can be used to hide the noise. This is the type of technique which is investigated here and is why we say the noise is being 'perceptually reduced'.

Two techniques which can be used to achieve this type of reduction are time domain harmonic scaling (TDHS) [1] and adaptive comb filtering [2]. These two techniques work quite differently and will be explained in the next section. Drawing on both the strengths and weaknesses of these two techniques a hybrid technique is described in the third section. It was applied to the frame rate noise of adaptive transform coding and these results are described in the fourth section.

Noise Reduction via TDHS and Adaptive Comb Filtering

TDHS can be viewed in the time domain as a pitch synchronous block interpolation or decimation scheme. For example, for 2:1 decimation two successive pitch periods are combined to form a single output period. The details of this process are very simple: each output point is a linear combination of two input points which were separated in time by one pitch period. The weights used in the combination are time varying and are used in such a way so as to insure continuity. The 2:1 decimation process using a triangular weighting function is schematically shown in Fig. 1a and 2:1 interpolation in Figure 1b.


(a) 2:1 DECIMATION

(b) 2:1 INTERPOLATION

Figure 1 TDHS compression and expansion operations.

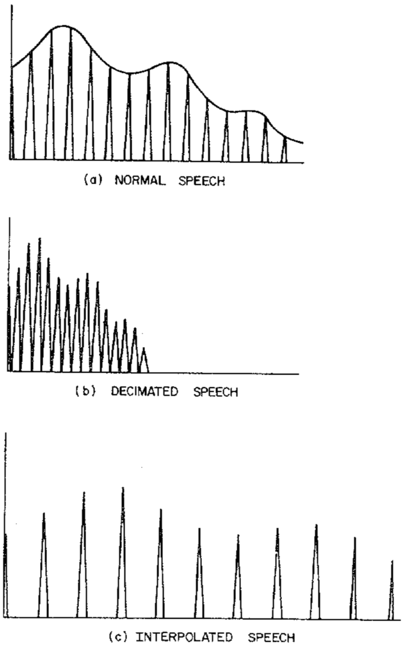
In the frequency domain the explanation is slightly more complex. We begin with the spectral characteristics of speech. In the frequency domain the pitch structure appears as "teeth" as shown schematically in Fig. 2a. Between the teeth are gaps. Block decimation consists of shifting the teeth downward in frequency, reducing the gaps as shown in Fig. 2b. Block interpolation means shifting them upwards, increasing the gaps as shown in Fig. 2c. The time domain window used determines the shape of the filter around each pitch tooth.

This is the basis for one clean up algorithm proposed here. First we will contract the spectrum, then expand it and in the process clean out the gaps. Consider the spectrum of periodically structured interference. An example would be a series of periodically spaced impulses. This spectrum would be a line spectrum with lines spaced f_0 apart where f_0 is the frequency of the impulses. If we superimposed the impulse noise on speech the two spectra would also add. But while the noise spectra would remain constant the speech spectrum would change with time as pitch changes.

Now consider the effect of the compression/expansion process on the combined spectrum. In general the fundamental frequency of the pitch would be different than that of the noise. Many of the noise lines would fall in the gaps. These are the frequencies we

[*] On leave from the Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa, Israel.

CH1610-5/81/0000-1089 $00.75 © 1981 IEEE


(a) NORMAL SPEECH

(b) DECIMATED SPEECH

(c) INTERPOLATED SPEECH

Figure 2 Effects of TDHS operations on the spectrum.

mostly hear because there is no speech content there. If the noise is low level then the noise lines falling on or near the pitch teeth will be masked by the speech. The compression/expansion process shifts the noise lines under the teeth as shown in Figure 3. In part "A" we show how each of the pitch teeth is filtered. All noise lines falling under a filter are shifted with the pitch tooth. If we carry out the compression process each pitch tooth is shifted to half its previous frequency but the increments between it and any nearby noise lines remain the same. For example if a pitch tooth and a noise line were 10 Hz apart before then they will still be 10 Hz apart after. Now if we play the compressed speech at the original sampling rate the result is natural sounding speech which only takes half as long. In the frequency domain this is equivalent to multiplying all frequencies by 2. This is the version of the spectrum shown in part "B". Note that the increments between any tooth and its noise line are now doubled. As a particular example consider the noise line marked "n". It is now closer to P_1 than P_2. In the expansion process the teeth are again filtered and now shifted upward. Line "n" actually falls in the transition regions of the filters for both P_1 and P_2, so it is shifted upward twice, once with each tooth, again maintaining its increments from P_1 and P_2. If we now play the speech at the original sampling rate this is equivalent to dividing all frequencies by 2. This is shown in part 3 and the two lines which evolved from "n" are labeled "n_1" and "n_2". As can be seen the overall tendency is to group noise lines close to the teeth leaving the gaps empty. In Fig. 4 a synthetic vowel spectra is shown before and after TDHS clean up from a periodic noise. The cleaned-up spectrum shows the same tendency to group the noise under the pitch teeth.

Adaptive comb filtering as described by Lim et al. in [2] is a one step operation that can be viewed as a stacking procedure. Figure 5 shows a method for implementing a 3 tap comb filter. Three successive pitch periods are weighted, each with a different weight, and then added together to give a single output period. In the frequency domain this has the effect of placing a filter around each pitch tooth and filtering out anything in the gaps. The method is adaptive because the filter length is updated when the pitch changes. In addition Lim [2] inserted zeros or deleted samples as necessary to make sure all the weighted pitch periods were the same length.

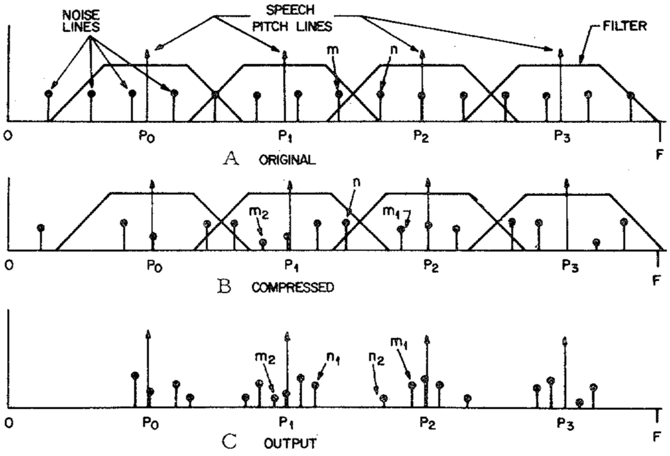


Figure 3 Effects of TDHS operations on interference.

The difference between TDHS noise reduction and comb filtering is that TDHS moves the noise in the gaps under the pitch tooth while comb filtering seeks to filter out the noise in the gaps. In addition, with TDHS it turns out that splitting the frequency components and their modulation to different frequencies also has the effect of reducing the noise energy. In studies made on several types of periodic noise we found the noise reduction in SNR for both techniques to be about equal. However, each one also contributed a noise or distortion of its own. TDHS mildly degrades the speech through the block decimation process in the contraction part of the algorithm, resulting in a slightly reverberant quality. Comb filtering introduces a discontinuity every time the pitch changes or there is a pitch error. In addition, if the pitch is slightly in error it has a tendency to attenuate the higher frequency pitch teeth instead of attenuating the noise in the gaps, whereas TDHS is more tolerant of small pitch errors [1,4].

This led us to synthesize a hybrid technique based on both of these techniques. If we first expanded and then contracted the spectrum via TDHS we get a lesser amount of clean-up but do not harm the speech as much. (We shall refer to the new technique as stretch-compress (S-C) TDHS.) The combined expansion and contraction operation is a time-varying adaptive comb filter. The difference is that with this technique there is guaranteed continuity when pitch changes which the previous comb filter did not have.

This new technique does not harm the speech the way either comb filtering or compress-stretch (C-S) TDHS does but gives comparable performance according to SNR measurements we made. In informal listening comparisons we felt that C-S TDHS removed more of the noise during speech but produced a most disagreeable noise pattern during non-speech segments and also slightly distorted the speech. Comb filtering and the new technique produced similar results except that comb filtering produced audible discontinuities in regions where the pitch changed. More details on S-C TDHS are given in the next section.

A Class of Windows for One Step Implementations

The expansion and compression processes were shown in Figure 1. In practice an S-C TDHS method which does not involve the additional storage necessitated by the sequential expansion and compression operations is desirable. To find such a method consider the combined expansion and compression equations:

Expansion

Let the expansion window $W_e(n)$ of duration $4P$ be defined by

$$W_e(n) = \begin{cases} \alpha(n) \quad , & n = 0,1, \ \ldots, 2P-1 \\ 1-\alpha(n-2P), & n = 2P, \ \ldots, 4P-1 \end{cases} \qquad (1)$$

where P is the pitch period and $\alpha(n)$ is the upward going half of

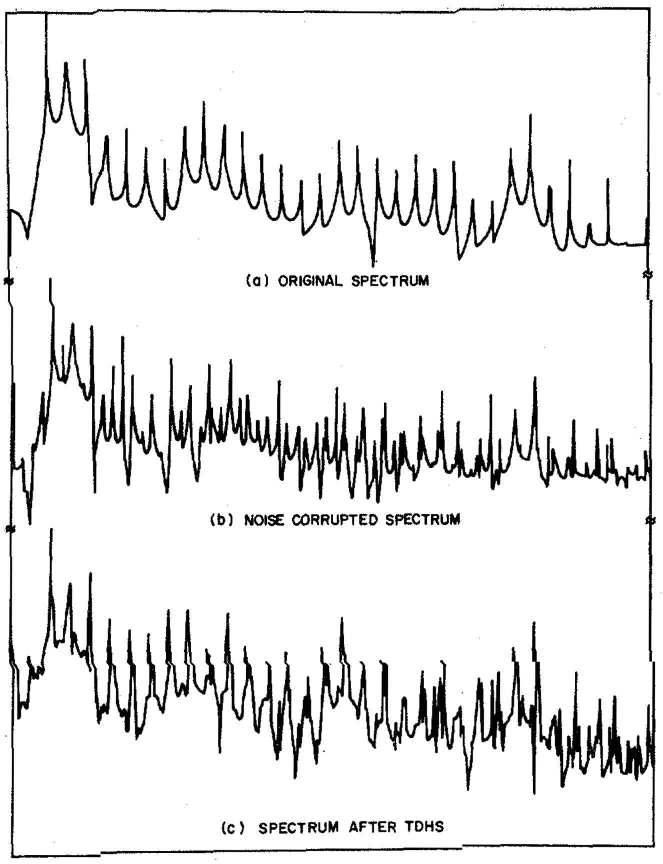


Figure 4 Example of TDHS on noise corrupted vowel.

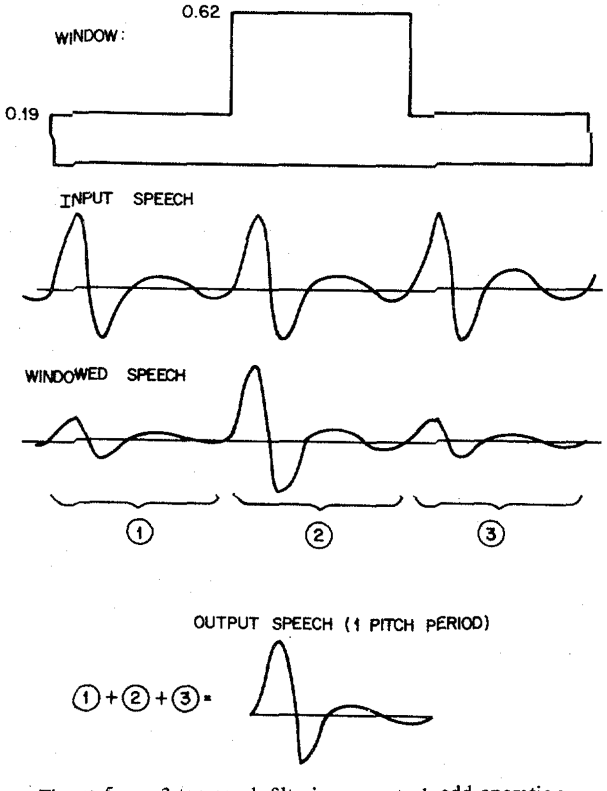


Figure 5 3 tap comb filtering as a stack-add operation.

the window. A segment of the expanded signal of duration $2P$ is then given by

$$Y(n) = [1-\alpha(n)]X(n) + \alpha(n)X(n-P) \quad n = 0,1,...,2P-1 \quad (2)$$

Compression

Let the compression window $W_c(n)$, of duration $2P$, be similarly defined by

$$W_c(n) = \begin{cases} \beta(n) \quad , & n = 0,1,...,P-1 \\ 1-\beta(n-P), & n = P, P+1,...,2P-1 \end{cases} \quad (3)$$

If the expanded signal segment $Y(n)$ in (2) is weighted by this compression window, a speech segment $Z(n)$ of duration P is obtained and is given by

$$Z(n) = [1-\beta(n)]\, Y(n) + \beta(n)Y(n+P) \quad (4)$$
$$n = 0,1, \cdots, P-1$$

substituting (2) into (4) we find

$$Z(n) = W_1(n)X(n-P) + W_2(n)\, X\,(n) + W_3(n)X(n+P) \quad (5)$$
$$n = 0,1, \cdots, P-1$$

where

$$W_1(n) = \alpha(n)\,[1-\beta(n)]$$
$$W_2(n) = [1-\alpha(n)]\,[1-\beta(n)] + \alpha(n+P)\,\beta(n)$$
$$W_3(n) = [1-\alpha(n+P)]\,\beta(n)$$
$$n = 0,1, \cdots, P-1 \quad (6)$$

It is common to use the same window shape for both the expansion and compression operation, i.e.

$$\beta(n) = \alpha(2n) \qquad n = 0,1, \cdots, P-1 \quad (7)$$

Substituting this relation in (6) results in

$$W_1(n) = \alpha(n)\,[1-\alpha(2n)]$$

$$W_2(n) = [1-\alpha(2n)]\,[1-\alpha(2n)] + \alpha(n+P)\,\alpha(2n)$$

$$W_3(n) = [1-\alpha(n+P)]\,\alpha(2n)$$
$$n = 0,1, \cdots, P-1 \quad (8)$$

Finally, we consider the use of a triangular window (A Hanning window is also an adequate window function [1]), for which

$$\alpha(n) = n/2P, \quad n = 0,1, \cdots, 2P-1 \quad (9)$$

Upon substitution of (9) into (8) the following combined window function used in our experiment results

$$W_1(n) = W_3(n) = \epsilon(n,P)$$
$$W_2(n) = 1 - 2\epsilon(n,P)$$
$$n = 0,1,..., P-1 \quad (10)$$

where

$$\epsilon(n,P) = n(P-n)/(2P^2) \quad (11)$$
$$n = 0,1,...,P-1$$

Note that

$$W_1(n) + W_2(n) + W_3(n) = 1 \quad (12)$$
$$n = 0,1, \cdots, P-1$$

which is an obvious requirement.

Rewriting (5) in terms of (10) results in

$$Z(n) = X(n) + \epsilon(n,P)\,[X(n-P) - 2X(n) + X(n+p)] \quad (13)$$
$$n = 0,1, \cdots, P-1$$

which shows that only one multiplication is needed (the multiplication by 2 can be carried out by a shift operation). For real-time implementation $\epsilon(n,P)$ can be precomputed for all possible values of P and stored. For the common range of pitch period duration and a sampling rate of 8 KHz only about 4096 storage locations are necessary.

In experimental trials with this window, equation (11) was modified with a multiplicative gain:

$$\epsilon(n,P) = Gn(P-n)/(2p^2) \qquad (14)$$

In informal listening we found better performance with a gain of G=2 than the pure TDHS window with G=1. The window described by (10) and (14) with G=2 is shown in Fig. 6.

Perceptually Removing ATC Frame Rate Noise

Adaptive Transform Coding (ATC) [3] encodes speech by selectively quantizing frequency components from a short-time spectrum. Successive blocks, typically about 32 msec long, are treated independently and often have different bit allocations for the spectrums. This can result in a discontinuity between successive blocks. If there is some overlap between blocks this lessens the low frequency noise but the high frequency components still remain and cause a burbling noise in the background.

In an earlier experiment [4] with ATC and TDHS to reduce the needed transmission bit rate it was noted that TDHS reduced the burble. This was the motivation for pursuing the clean up properties .of TDHS. In this experiment ATC was simulated using the algorithm of Cox and Crochiere [3] at bit rates of 7.2, 9.6, 12 and 16 Kb/s. The new clean-up algorithm was applied to each of the output signals. There were six input sentences. An experiment tape was generated in which all possible conditions were compared in both an A-B and B-A comparison. The 9 conditions were the original, the 4 coder outputs and the 4 clean-up versions. This resulted in 72 comparisons. The tape was played for a dozen subjects who were asked to indicate their preference for each comparison.

The overall results are shown in Figure 7 and Table I gives the test data. Each condition could receive a maximum of 192 votes overall and 24 in comparison with any one other condition. As expected the original received the highest percentage, 96%. The important result was that at the lower bit rates of 7.2, 9.6 and 12 Kb/s the clean-up algorithm subjectively improved the quality of the coder output. The 7.2 Kb/s cleaned-up version was rated equivalent to the 9.6 Kb/s original ATC version and the 12 Kb/s cleaned-up version was rated slightly less than the 16 Kb/s original ATC version. At 16 kb/s the cleaned-up speech was rated equal to the coded speech. Since at 16 kb/s the coded speech has no noticeable frame rate noise, this shows that the clean-up technique does not degrade the speech. These results indicate that this clean-up technique is very effective at low bit rates. This is significant because these are the bit rates of interest for transform coding.

Conclusions

A new speech enhancement procedure based on TDHS has been introduced. It can be viewed as a form of time varying adaptive comb filtering. However, unlike previous comb filtering techniques there is no discontinuity produced when pitch changes and it is less susceptible to small errors in pitch. These properties derive from its TDHS origins. Its implementation is straightforward and its usefulness has been demonstrated in reducing noise from ATC coded speech.

	ATC				ATC-CL			
	7.2	9.6	12	16	7.2	9.6	12	16
Original	100	96	100	96	100	96	100	83
ATC 7.2	-	33	8	0	17	8	12	0
ATC 9.6	67	-	37	12	67	25	17	17
ATC 12	92	63	-	17	54	46	25	29
ATC 16	1Q0	88	83	-	88	83	50	50
ATC-CL 7.2	83	33	46	12	-	42	17	12
ATC-CL 9.6	92	75	54	17	58	-	21	21
ATC-CL 12	88	83	75	50	83	79	-	42
ATC-CL 16	100	83	71	50	88	79	58	-

Table I Coder vs. Coder Comparisons

References

[1] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals", IEEE Trans. on ASSP, vol. ASSP-27, pp. 121-133, April 1979.

[2] J. S. Lim, A. V. Oppenheim, and L. D. Braido, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-26, pp. 354-358, August 1978.

[3] R. V. Cox and R. E. Crochiere, "Real-time simulation of adaptive transform coding", IEEE Trans. on ASSP, to be published.

[4] D. Malah, R. E. Crochiere, and R. V. Cox, "Performance of transform and sub-band coding systems combined with harmonic scaling of speech", IEEE Trans. on ASSP, to be published.

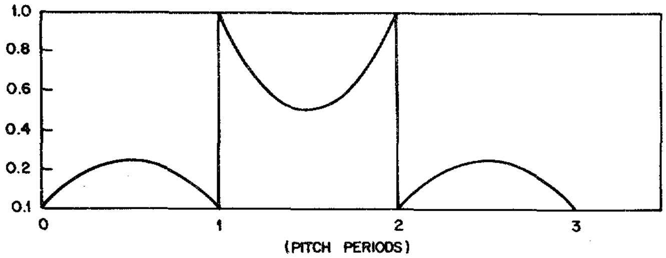


Figure 6 S-C TDHS window used.

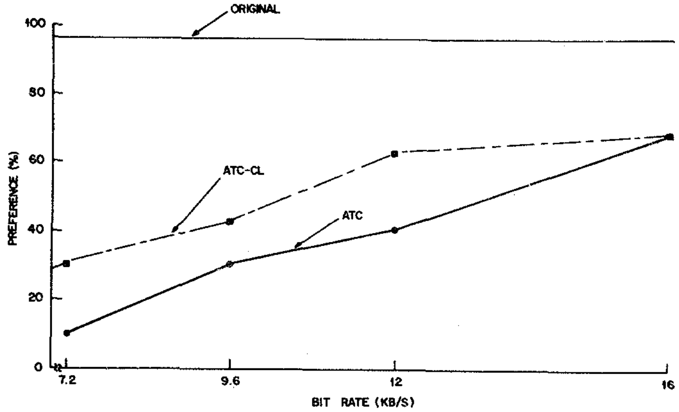


Figure 7 Results of subjective test.