ENHANCEMENT OF SPEECH CORRUPTED BY ACOUSTIC NOISE*

M. Berouti, R. Schwartz, and J. Makhoul

Bolt Beranek and Newman Inc.
Cambridge, Mass.

ABSTRACT

This paper describes a method for enhancing speech corrupted by broadband noise. The method is based on the spectral noise subtraction method. The original method entails subtracting an estimate of the noise power spectrum from the speech power spectrum, setting negative differences to zero, recombining the new power spectrum with the original phase, and then reconstructing the time waveform. While this method reduces the broadband noise, it also usually introduces an annoying "musical noise". We have devised a method that eliminates this "musical noise" while further reducing the background noise. The method consists in subtracting an overestimate of the noise power spectrum, and preventing the resultant spectral components from going below a preset minimum level (spectral floor). The method can automatically adapt to a wide range of signal-to-noise ratios, as long as a reasonable estimate of the noise spectrum can be obtained. Extensive listening tests were performed to determine the quality and intelligibility of speech enhanced by our method. Listeners unanimously preferred the quality of the processed speech. Also, for an input signal-to-noise ratio of 5 dB, there was no loss of intelligibility associated with the enhancement technique.

1. INTRODUCTION

We report on our work to enhance the quality of speech degraded by additive white noise. Our goal is to improve the listenability of the speech signal by decreasing the background noise, without affecting the intelligibility of the speech. The noise is at such levels that the speech is essentially unintelligible out of context. We use the average segmental signal-to-noise ratio (SNR) to measure the noise level of the noise-corrupted speech signal. We found that sentences with a SNR in the range -5 to +5 dB have an intelligibility score in the range 20 to 80%. There is strong correlation between the intelligibility of a sentence and the SNR, but intelligibility also depends on the speaker, on context, and on the phonetic content.

After an initial investigation of several methods of speech enhancement, we concluded that the method of spectral noise subtraction is more effective than others. In this paper we discuss our implementation of that method, which differs from that reported by others in two major ways: first, we subtract a factor (α) times the noise spectrum, where α is a number greater than unity and varies from frame to frame. Second, we prevent the spectral components of the processed signal from going below a certain lower bound which we call the spectral floor. We express the spectral floor as a fraction β, of the original noise power spectrum $P_n(w)$.

2. BASIC METHOD

The basic principle of spectral noise subtraction appears in the literature in various implementations [1-4]. Basically, most methods of speech enhancement have in common the assumption that the power spectrum of a signal corrupted by uncorrelated noise is equal to the sum of the signal spectrum and the noise spectrum. The preceding statement is true only in the statistical sense. However, taking this assumption as a reasonable approximation for short-term (25 ms) spectra, its application leads to a simple noise subtraction method. Initially, the method we implemented consisted in computing the power spectrum of each windowed segment of speech and subtracting from it an estimate of the noise power spectrum. The estimate of the noise is formed during periods of "silence". The original phase of the DFT of the input signal is retained for resynthesis. Thus, the enhancement algorithm consists of a straightforward implementation of the following relationship:

$$\text{let } D(w) = P_s(w) - P_n(w)$$

$$P_s'(w) = \begin{cases} D(w), & \text{if } D(w) > 0 \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

where $P_s'(w)$ is the modified signal spectrum, $P_s(w)$ is the spectrum of the input noise-corrupted speech, and $P_n(w)$ is the smoothed estimate of the noise spectrum. $P_n(w)$ is obtained by a two-step process: First we average the noise spectra from several frames of "silence". Second, we smooth in frequency this average noise spectrum. For the specific case of white noise, the smoothed estimate of the noise spectrum is flat. The enhanced speech signal is obtained from both $P_s'(w)$ and the original phase by an inverse Fourier transform:

$$s'(t) = F^{-1}\{\sqrt{P_s'(w)}\ e^{j\theta(w)}\} \qquad (2)$$

where $\theta(w)$ is the phase function of the DFT of the input speech. Since the assumption of uncorrelated signal and noise is not strictly valid for short-term spectra, some of the components of the processed spectrum, $P_s'(w)$, may be negative. These negative values are set to zero as shown in (1).

* An earlier version of this paper was presented at the ARPA Network Speech Compression (NSC) Group meeting, Cambridge, MA, May 1978, in a special session on speech enhancement.

CH1379-7/79/0000-0208$00.75 © 1979 IEEE

A major problem with the above implementation of the spectral noise subtraction method has been that a "new" noise appears in the processed speech signal. The new noise is variously described as ringing, warbling, of tonal quality, or "doodly-doos". We shall henceforth refer to it as the "musical noise". Also, though the noise is reduced, there is still considerable broadband noise remaining in the processed speech.

3. NATURE OF THE PROBLEM

To explain the nature of the musical noise, one must realize that peaks and valleys exist in the short-term power spectrum of white noise; their frequency locations for one frame are random and they vary randomly in frequency and amplitude from frame to frame. When we subtract the smoothed estimate of the noise spectrum from the actual noise spectrum, all spectral peaks are shifted down while the valleys (points lower than the estimate) are set to zero (minus infinity on a logarithmic scale). Thus, after subtraction there remain peaks in the noise spectrum. Of those remaining peaks, the wider ones are perceived as time varying broadband noise. The narrower peaks, which are relatively large spectral excursions because of the deep valleys that define them, are perceived as time varying tones which we refer to as musical noise.

4. PROPOSED SOLUTION

Our modification to the noise subtraction method consists in minimizing the perception of the narrow spectral peaks by decreasing the spectral excursions. This is done by changing the algorithm in (1) to the following:

$$\text{let } D(w) = P_s(w) - \alpha P_n(w)$$

$$P_s'(w) = \begin{cases} D(w), & \text{if } D(w) > \beta P_n(w) \\ \beta P_n(w), & \text{otherwise} \end{cases} \qquad (3)$$

$$\text{with } \alpha \geq 1, \text{ and } 0 < \beta << 1$$

where α is the subtraction factor and β is the spectral floor parameter. The modified method is

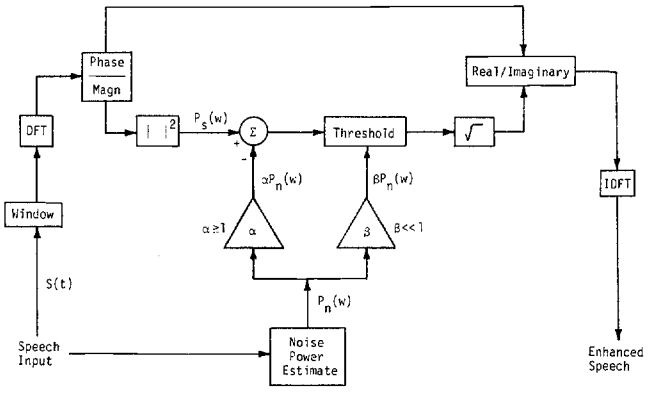

Fig. 1 Modified spectral noise subtraction method, with spectral floor.

shown in Fig. 1. Note that (3) is identical to (1) for $\alpha=1$ and $\beta=0$.

From (3) it can be seen that the goal of reducing the spectral noise peaks can be achieved with $\alpha>1$. For $\alpha>1$ the remnants of the noise peaks will be lower relative to the case with $\alpha=1$. Also, with $\alpha>1$ the subtraction can remove all of the broadband noise by eliminating most of the wide peaks. However, this by itself is not sufficient, because the deep valleys surrounding the narrow peaks remain in the noise spectrum and, therefore, the excursion of noise peaks remains large. The second part of our modification consists of "filling-in" the valleys. This is done in (3) by means of the spectral floor, $\beta P_n(w)$: The spectral components of $P_s'(w)$ are prevented from descending below the lower bound $\beta P_n(w)$. For $\beta>0$, the valleys between peaks are not as deep as for the case $\beta=0$. Thus, the spectral excursion of noise peaks is not as large, which reduces the amount of the musical noise perceived. Another way to interpret the above is to realize that, for $\beta>0$, the remnants of noise peaks are now "masked" by neighboring spectral components of comparable magnitude. These neighboring components in fact are broadband noise reinserted in the spectrum by the spectral floor $\beta P_n(w)$. Indeed, speech processed by the modified method has less musical noise than speech processed by (1). We note here that for $\beta<<1$ the added broadband noise level is also much lower than that perceived in speech processed by (1).

In order to be able to refer to the "broadband noise reduction" achieved by the method, we have conveniently expressed the spectral floor as a fraction of the original noise power spectrum. Thus, when the spectral floor effectively masks the musical noise, and when all that can be perceived is broadband noise, then the noise attenuation is given by β. For instance, for $\beta=0.01$, there is a 20 dB attenuation of the broadband noise.

Various combinations of α and β give rise to a trade-off between the amount of remaining broadband noise and the level of the perceived musical noise. For β large, the spectral floor is high, and very little, if any, musical noise is audible, while with β small, the broadband noise is greatly reduced, but the musical noise becomes quite annoying. Similarly, we have found that, for a fixed value of β, increasing the value of α reduces both the broadband noise and the musical noise. However, if α is too large the spectral distortion caused by the subtraction in (3) becomes excessive and the speech intelligibility may suffer.

In practice, we have found that at SNR=0 dB, a value of α in the range 3 to 6 is adequate, with β in the range 0.005 to 0.1. A large value of α, such as 5, should not be alarming. This is equivalent to assuming that the noise power to be subtracted is about 7 dB higher than the smoothed estimate. This "inflation" factor represents the fact that, at each frame, the variance of the spectral components of the noise is equal to the noise power itself. Hence, one must subtract more than the expected value of the noise spectrum (the smoothed estimate) in order to make sure that most of the noise peaks have been removed.

In order to reduce the speech distortion caused by large values of α, we decided to let α vary from frame to frame within the same sentence. To understand the rationale behind doing so, consider the graph of Fig. 2. The dotted line in the figure shows a plot of the value of α used in an experiment where several sentences at different SNR were processed. In the experiment, α was constant for each utterance. At the completion of the experiment, we noticed that the optimal value of α, as determined empirically for best noise reduction with the least amount of musical noise,

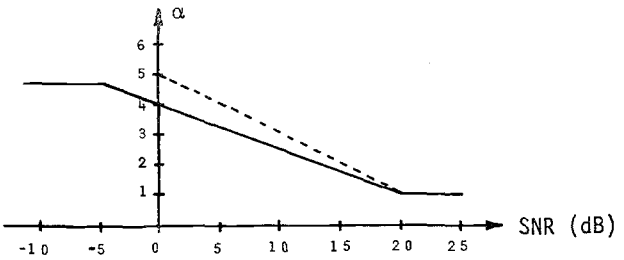

Fig. 2 Value of the subtraction factor α versus the SNR.

is smaller for higher SNR inputs. We then decided that α could vary not only across sentences with different SNR but also across frames of the same sentence. The reason for allowing α to vary within a sentence is that the segmental SNR varies from frame to frame in proportion to signal energy because the noise level is constant. After extensive experimentation, we found that α should vary within a sentence according to the solid line in Fig. 2, with α=1 for SNR$\geq$20 dB. Also, we prevent any further increase in α for SNR<-5 dB. The slope of the line in Fig. 2 is determined by specifying the value of the parameter α at SNR=0 dB. The SNR is estimated at each frame from knowledge of the noise spectral estimate and the energy of the input speech. At each frame, the actual value of α used in (3) is given by:

$$\alpha = \alpha_0 - (\mathrm{SNR})/s \quad \text{for} \quad -5 \leq \mathrm{SNR} \leq 20 \tag{4}$$

where α_0 is the desired value of α at SNR=0 dB, SNR is the estimated segmental signal-to-noise ratio and 1/s is the slope of the line in Fig. 2. (For example, for α_0=4, s=20/3.) We found that using a variable subtraction reduces the speech distortion somewhat. If the slope (1/s) is too large, however, the temporal dynamic range of the speech becomes too large.

To summarize, there are several qualitative aspects of the processed speech that can be controlled. These are: the level of the remaining broadband noise, the level of the musical noise, and the amount of speech distortion. These three effects are controlled mainly by the parameters α_0 and β.

5. OTHER RELATED PARAMETERS

Aside from the parameters α and β discussed above, we investigated several other parameters. These are:

a) the exponent of the power spectrum of the input (so far assumed to be 1),
b) The normalization factor needed for output level adjustment,
c) the frame size,
d) the amount of overlap between frames,
e) the FFT order.

All of the above parameters interact with each other and with α and β. We shall now discuss each parameter individually.

Exponent of the Power Spectrum

We investigated raising the power spectrum of the input to some power γ before the subtraction. In this case, (3) becomes:

$$\text{let } D(w) = G[P_s^{\gamma}(w) - \alpha P_n^{\gamma}(w)]$$

$$P_s'(w) = \begin{cases} D^{1/\gamma}(w), & \text{if } D^{1/\gamma}(w) > \beta P_n(w) \\ \beta P_n(w), & \text{otherwise} \end{cases} \tag{5}$$

$$\text{with } \alpha \geq 1, \text{ and } 0 < \beta << 1$$

where G is the normalization factor to be discussed later. Note that (5) is identical to (3) for γ=1 and G=1. Equation (5) is implemented by means of the same algorithm illustrated in Fig. 1, except that all symbols $P_x(w)$ are replaced by $P_x^{\gamma}(w)$ and the gain G follows the subtraction in Fig. 1 and precedes the thresholding. For a fixed value of α_0, the subtraction in (5) with a value of γ<1 results in a greater amount of spectral change than for the case γ=1. We note here that Boll [2,3] uses γ=0.5, with α=1 and β=0, whereas Suzuki et al. [1] and Curtis and Niederjohn [4] use γ=1.

Normalization Factor

The next parameter to consider is a normalization factor to scale the processed signal. Our initial experiments were all done with γ=1 and we found no need for such normalization. However, for γ<1 the subtraction affects the spectrum more drastically than for the case γ=1. Therefore, for lower γ, the processed output had an extremely low level, which prevented us from comparing sentences that were processed with different values of γ. Our initial approach to normalization was to force the energy of the processed signal at each frame to be equal to the difference between the input energy and the estimated noise energy. Once again, we were relying on the assumption that the signal and the noise are uncorrelated. This approach required that the normalization factor change drastically from frame to frame, which led to severe problems, especially in low energy frames. In our final approach, we corrected the problem by keeping the normalization factor constant over most of the sentence. We accomplished this by starting with a high initial value for the normalization factor, and updating its value at high energy frames only. The update takes place only if the newly derived factor is smaller than the previous one. In practice, we compute $A=(1/\gamma)(P_s-P_n)/P_d$, for $P_s \geq 2P_n$, where P_s, P_n and P_d are the estimated power of the signal, power of the noise, and power of the signal processed without the gain. If the value of A obtained is less than the previous value, we update the value of the normalization factor $G=A^{\gamma}$. Also, G is not allowed to be less than 1.0. The effect of the normalization is to keep the average level of the processed speech independent of the power γ used. Finally, we note that normalization takes place after the subtraction, but before the application of the spectral floor constraint. In this fashion, it is still possible to relate the spectral floor to the original input noise power by means of the constant β, irrespective of which power γ was used for the processing in (5). Thus, the perceived remaining broadband noise is determined only by $\beta P_n(w)$.

Frame Size

The frame size had been set to 25 ms throughout the initial phase of our work. We have found that using an analysis frame shorter than 20 ms results in roughness, while increasing the frame size decreases the musical noise considerably. However, if the frame is too long, slurring results.

Window Overlap

Associated with the frame size is the amount of overlap between consecutive frames. We have used the Tukey window (flat in its middle range and with cosine tapering at each end) in order to overlap and add adjacent segments of processed speech. The overlap is necessary to prevent discontinuities at frame boundaries. The amount of overlap is usually taken to be 10% of the frame size. However, for larger frames, 10% may be excessive and might cause slurring of the signal.

FFT Order

The third window-related parameter is the order of the FFT. In general, enough zeros are appended at one end of the windowed data prior to obtaining the DFT, such that the total number of points is a power of 2 and, thus, an FFT routine can be used. However, processing in the frequency domain causes the non-zero valued data to extend out of its original time-domain range into the added zeros. If the added-zero region is not long enough, time-domain aliasing might occur. Thus we needed to investigate adding more zeros and using a higher order FFT.

6. EXPERIMENTS AND RESULTS

The discussions in Sections 3 and 4 shed some light on the effect that each parameter has on the quality of the processed speech. We performed several experiments to understand further how all these parameters interact. We were mainly interested in finding an optimal range of values for α_0 and β. As mentioned earlier, these two parameters give us direct control of the three major qualitative aspects of processed speech: remaining broadband noise, musical noise, and speech distortion. Clearly, we desire values of α_0 and β that would minimize those three effects. However, the effects of the parameters α_0 and β on the quality of the processed speech are intimately related to the input SNR, the power γ, and the window-related parameters.

Throughout our experiments, we considered inputs with SNR in the range -5 to +5 dB and used values of γ=0.25, 0.5, and 1. We have experimented with several frame sizes (15 to 60 ms), different amounts of overlap between frames, and different FFT orders.

Through extensive experimentation we determined the range of values for each of the parameters of the algorithm. The ranges given below are meant to be guidelines rather than final "optimal" values. Optimality is a subjective choice and depends on the user's preference. Below we give some of the conclusions we reached:

- Frame size: The frame size should be between 25 and 35 ms.
- Overlap: The overlap between frames should be on the order of 2 to 2.5 ms.
- FFT order: Our investigations did not show that time-domain aliasing was an important issue. Therefore, the minimum FFT order corresponding to a given frame size is adequate, with no noticeable improvement in going to a higher order. The same was reported earlier by Boll [2].
- Exponent of the power spectrum: Of the three values of γ we tried, γ=1 was found to yield better output quality, in general.
- Subtraction factor: for γ=1, an optimal range for α_0 is 3 to 6 (for γ=.5, α_0 should be in the range 2 to 2.2). The slope in (4) (or Fig. 2) is set such that α=1 for SNR$\geq$20 dB, and $\alpha=\alpha_0$ at SNR=0 dB.
- Spectral floor: The spectral floor depends on the average segmental SNR of the input, i.e., the noise level. For high noise levels (SNR=-5 dB) β should be in the range 0.02 to 0.06, and for lower noise levels (SNR=0 or +5 dB) β should be in the range 0.005 to 0.02.

Towards the end of our research we performed a formal listening test to assess the quality and intelligibility of the enhanced speech. The input speech varied in SNR from -5 to +5 dB. The processing was done using parameter values as given by the above guidelines. Subjects unanimously preferred the quality of the enhanced speech to that of the unprocessed signal. In addition, at input SNR=+5 dB, using the values α_0=3, β=.005, γ=1, and a 32 ms frame size, the intelligibility of the enhanced speech was the same as that of the unprocessed signal. For lower SNR's, the intelligibility of the speech decreased somewhat. Prior to performing the formal intelligibility test, our algorithm had been tuned for optimal quality, i.e., maximum noise reduction, without accurate knowledge of the effect of the method on speech intelligibility. We believe that it may be possible to maintain the same intelligibility while improving the listenability of the speech by further tuning the parameters of the system (mainly α_0 and β). The actual parameter values used in a specific situation depend on one's purpose in using the enhancement algorithm. In some applications a slight loss of intelligibility may be tolerable, provided the listenability of the speech is greatly improved. In other applications a loss in intelligibility may not be acceptable.

7. CONCLUSIONS

To conclude, the main differences between the basic spectral subtraction method and our implementation is that we subtract an overestimate of the noise spectrum and prevent the resultant spectral components from going below a spectral floor. Our implementation of the spectral noise subtraction method affords a great reduction in the background noise with very little effect on the intelligibility of the speech. Formal tests have shown that, at SNR=+5 dB, the intelligibility of the enhanced speech is the same as that of the unprocessed signal.

ACKNOWLEDGMENTS

The authors wish to thank A.W.F. Huggins for his contributions to this research. This work was sponsored by the Department of Defense.

REFERENCES

1. H. Suzuki, J. Igarashi, and Y. Ishii, "Extraction of Speech in Noise by Digital Filtering," J. Acoust. Soc. of Japan, Vol. 33, No. 8, Aug. 1977, pp. 405-411.
2. S. Boll, "Suppression of Noise in Speech Using the SABER Method," ICASSP, April 1978, pp. 606-609.
3. S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," submitted, IEEE Trans. on Acoustics, Speech and Signal Processing.
4. R.A. Curtis, R.J. Niederjohn, "An Investigation of Several Frequency-Domain Methods for Enhancing the Intelligibility of Speech in Wideband Random Noise," ICASSP, April 1978, pp. 602-605.