M.Tech. Credit Seminar Report, Electronic Systems group, EE Dept., IIT Bombay, Submitted November 2003

SPEECH ENHANCEMENT BY SPECTRAL SUBSTRACTION

Umesh Chandra Naik(03307424)
Supervisor: Prof. P.C. Pandey

Abstract

Considerable attention has been focused on the problem of enhancement of speech in acoustic background noise over last several years, motivated by application in speech transmission, coding and recognition. The spectral substraction method is a single-channel speech enhancement method which is popular due to its simplicity and computational efficiency. In this method, a noise estimate is substracted from the noisy speech spectrum. The noise spectrum is estimated and updated during the speech silence period. The spectral substraction technique performs well as a pre-processor noise reduction technique for digital voice processors. In our work we first discuss the need for speech enhancement, its applications and available different approaches. It is followed by classification of single channel enhancement techniques and brief overview of each method is given. The basic spectral substraction method and a modified version that minimises the shortcomings of the basic methods have been discussed in detail. Conclusions are drawn on the performance of spectral substraction method. Finally the limitations of the present spectral substraction method have been identified and suggestion for further improvement have been put forward.

1.INTRODUCTION

In many speech communication settings, the presence of background interference causes the quality and/or intelligibility of speech to degrade. A noisy environment can induce listener fatique and reduce the listener's ability to understand what is said. In addition to interpersonal communication, the quality of speech can also be influenced in data conversion, transmission or reproduction. The purpose of many enhancement algorithm is to reduce the background noise, improve speech-quality or suppress channel or speaker interference. In our work we study one such enhancement technique to enhance the quality of speech in presence of additive broadband acoustic noise.

Broadly the enhancement techniques can be classified as single channel, dual channel or multi-channel enhancement techniques. Single channel enhancement techniques apply to situations in which only one acquisition channel is available, for example, voice telephone, radio communication, mobile telephony. In dual channel enhancement techniques the acoustic sound waves arrive at each sensor. Two different approaches are used for dual-channel algorithm. In the first, a primary channel contains speech with additive noise and a second channel contains sample noise correlated with the noise in the primary channel. Normally, an acoustic barrier exists between sensors to ensure that no speech leaks into the noise interference channel. In the second approach, no acoustic barrier exists, so the enhancement algorithm must address the issue of cross-talk.

There are four broad classes of enhancement that differ substantially in the general approaches taken [3]. The first class concentrates on the short -term spectral domain. These techniques suppress noise by substracting an estimated noise bias found during non-speech activity in single microphone case or from reference microphone in dual-channel setting

The second class of enhancement technique is based on speech modelling using iterative methods. These systems focus on estimating model parameters that characterises the speech signal, followed by re-synthesis of the noise-free signal based on non-casual Wiener filtering. These enhancement techniques estimate speech parameters in noise based on auto-regressive, constrained autoregressive and autoregressive moving average models. This class of enhancement technique requires a priori knowledge of noise and speech statistics [3].

The third class of system is based on adaptive noise cancellation (ANC). Besides the degraded signal, a reference signal that is uncorrelated with the actual signal but correlated with the additive noise is needed for this method. With the help of this reference signal, an estimate of the noise can be made and substracted from the degraded speech. But since we have only single acquisition channel, we have to generate the reference signal from the available degraded speech.

The last area of enhancement is based on the periodicity of voiced speech. The periodicity of a waveform in time domain manifests itself in the frequency domain as harmonics with the fundamental frequency corresponding to the period of waveform. Hence the energy of the speech signal is concentrated only in narrow bands of frequency, whereas the interfering signals in general have energy over the entire frequency band. Using this concept, the comb filtering technique approach pass the harmonics of speech but rejects the frequency component between harmonics using a filter.

In our work single channel speech enhancement is studied because in most application only noisy speech is available. For such type of applications, short-time spectral domain methods are most suitable.

2. SPECTRAL SUBSTRACTION METHOD

In the earlier chapter we have studied the various methods available to deal with the problem of single channel enhancement of noisy speech. Among all these methods spectral substraction method is the most popular choice when one has to eliminate the background stationary noise. One drawback of the basic spectral algorithm suggested by Boll [1], that it generates musical noise. In this section we will first discuss the basic spectral substraction method and in the next section a modified spectral substraction method is studied in detail, which has been suggested by Berouti [2].

2.1 Overview of Basic Spectral Substraction Method

The basic assumption of the method is treating the noise as uncorrelated additive noise, which is true in case of background noise. This allows us to treat the power spectrum of the degraded speech as equal to the sum of the signal power spectrum and noise power spectrum [1].

$$y(n) = s(n) + d(n) \tag{1}$$

where s(n) represents the actual speech signal, $d(n)$ is uncorrelated additive noise and $y(n)$ represents the degraded speech signal.

Another assumption that we make is that of assuming $s(n)$ and $d(n)$ to be stationary signal as processing is done on a short time basis. Also, this method exploits the assumption that, for human perception the short time spectral amplitude is more important than phase for intelligibility and quality. This assumption has been shown to be true, by many works, one such work is of Lim and Wang [4], wherein they observed that using an actual phase rather than the degraded speech phase doesn't improve the quality of enhanced speech. Considering the signal model and taking Fourier transform of equation (1) we get

$$Y(w) = S(w) + D(w) \tag{2}$$

But as processing is carried out on a short-time basis, denoting the corresponding windowed signal by $y_w(n)$, $s_w(n)$, $d_w(n)$ respectively we have

$$y_w(n) = s_w(n) + d_w(n) \tag{3}$$

and its short-time Fourier transform is given by

$$|Y_w(w)|^2 = |S_w(w)|^2 + |D_w(w)|^2 + S_w(w).D_w^*(w) + D_w(w).S_w^*(w) \tag{4}$$

where $D_w^*(w)$, $S_w^*(w)$ represents complex conjugate of $D_w(w)$ and $S_w(w)$.

The function $|S_w(w)|^2$ is referred as the short-time energy spectrum of speech. For speech enhancement based on the short-time spectral amplitude, the objective is to obtain an estimate $\left|\hat{S}_w(w)\right|^2$ of $|S_w(w)|^2$ and from this, an estimate $\hat{s}_w(n)$ of $s_w(n)$. In the spectral substraction method [1], the terms $|D_w(w)|^2$, $S_w(w).D_w^*(w)$, $D_w(w).S_w^*(w)$ which can't be obtained directly are approximated by $E\left[|D_w(w)|^2\right]$, $E\left[S_w(w).D_w^*(w)\right]$ and $E\left[D_w(w).S_w^*(w)\right]$ where $E\left[.\right]$ denotes the ensemble average. Since we have assumed the noise to be uncorrelated with the signal, $E\left[S_w(w).D_w^*(w)\right]$ and $E\left[D_w(w).S_w^*(w)\right]$ are zero and estimate $\left|\hat{S}_w(w)\right|^2$ of $|S_w(w)|^2$ are given by

$$|\hat{S}_w(w)|^2 = |Y_w(w)|^2 - E[|D_w(w)|^2] \tag{5}$$

where $E[|D_w(w)|^2]$is obtained by either from the assumed known properties of $d(n)$ or by an actual measurement from background noise in the intervals of silence. For the second option speech silence detection algorithms are required.

Since the estimate $\left|\hat{S}_w(w)\right|^2$ can become negative due to over-estimation of noise, it is made equal to zero in those cases. This is sometimes known as half-wave rectification.

Let $\left|\hat{D}_w(w)\right|^2$=$E\left[|D_w(w)|^2\right]$ be the estimate of noise. Hence the complete power spectrum substraction method is mathematically given by

$$|\hat{S}_w(w)|^2 = |Y_w(w)|^2 - |\hat{D}_w(w)|^2, \quad if \quad |\hat{S}_w(w)|^2 > 0 \tag{6}$$

$$=0 \text{ otherwise}$$

Going by our assumption that short time phase is relatively unimportant,we approximate $\angle S_w(w)$, the phase of S_w, by $\angle Y_w(w)$ so that

$$\hat{S}_w = |\hat{S}_w|.exp[j\angle Y_w(w)] \tag{7}$$

$$and\ \hat{s}_w(n) = F^{-1}\{\hat{s}_w(w)\} \tag{8}$$

2.2 Drawback of Basic Spectral Substraction Method

The major problem of the basic method is that the algorithm may itself introduces a synthetis noise called musical noise. To explain its existence one must note the fact that we substract a smoothed estimate of noise spectrum whereas the short time power spectrum of actual white noise includes peaks and valley [2]. Thus after substraction there remain peaks in the noise spectrum. Of those remaining peaks the narrower peaks which are large spectral excursion because of the deep valleys that defines them, are perceived as time varying notes which we refer to as musical noise. The wider ones are perceived as time varying broadband noise.

Besides the above problem of musical noise, though the noise is reduced, considerable broadband noise still remain in the processed speech due to inherent mismatch between the noise estimate and actual noise content. This is referred as residual noise

3. MODIFIED SPECTRAL SUBSTRACTION METHOD

In order to minimise the presence of residual noise and the musical noise in the processed speech, we need to modify the spectral substraction method. Many modified spectral substraction methods have been suggested and implemented over last few years. Here we discuss one such method suggested by Berouti [2].

3.1 Algorithm

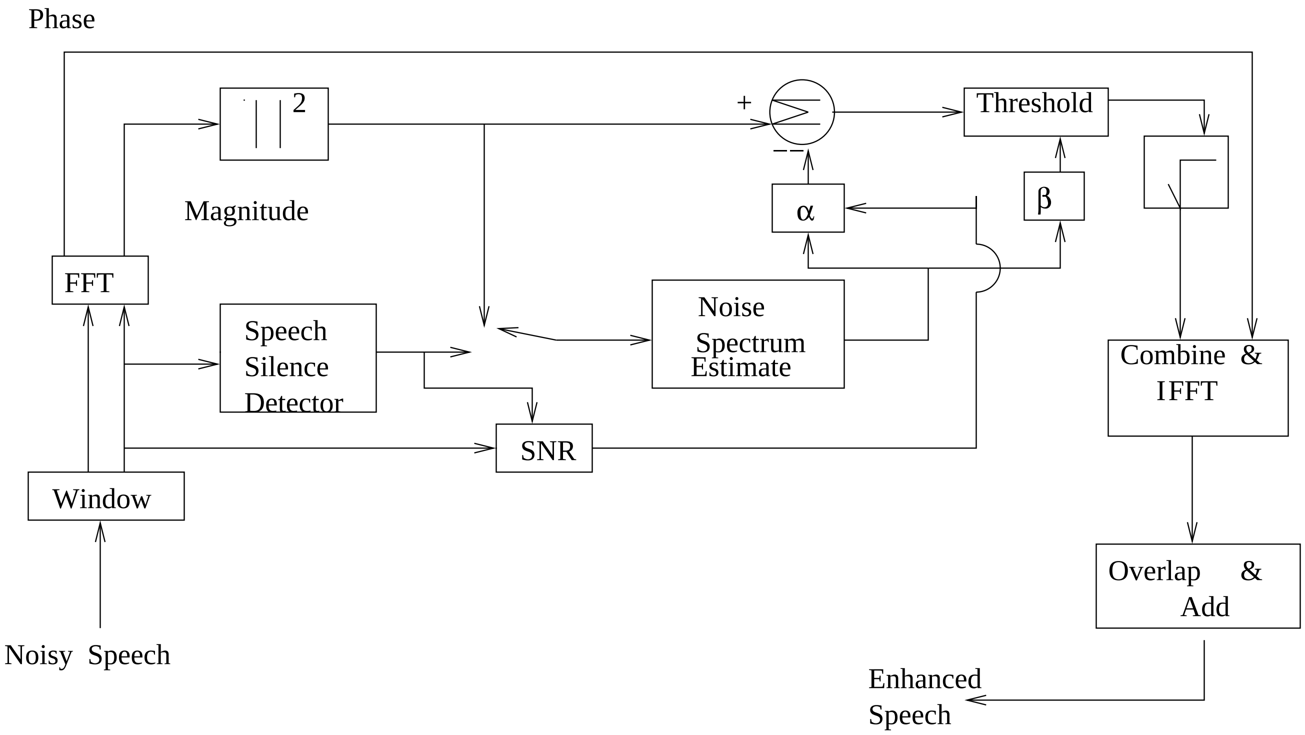

Figure 1: Block diagram of modified spectral substraction method (adapted from [5]).

Modification made to the original method are- substracting an overestimate of noise power spectrum and preventing the resultant spectrum from going below a preset minimum level (spectral floor).

Thus the algorithm of basic method given by (6) is modifiedied as

$$\begin{aligned} |\hat{S}_w(w)|^2 &= |Y_w(w)|^2 - \alpha|\hat{D}_w(w)|^2 \;\; if \;\; |\hat{S}_w(w)|^2 > \beta|\hat{D}_w(w)|^2 \\ &= \beta|\hat{D}_w(w)|^2 \quad otherwise \qquad (9) \\ \text{with } \alpha &\geq 1 \quad and \quad 0 \leq \beta = 1 \end{aligned}$$

where α is the substraction factor and β is the spectral floor parameter.

It can be seen from (9) that spectral noise peak will be lower with $\alpha = 1$ i.e, the basic method approach. Also when $\alpha > 1$, the substraction can remove all of the broadband noise eliminating most of the wide peaks. But deep valleys surrounding the narrow spectrum will remain in enhanced speech. This can be reduced by filling in the valleys. This objective is achieved by the spectral floor $\beta \left|\hat{D}_w(w)\right|^2$. For $\beta > 0$, the valleys between the peaks are not as deep as for the case β=0. Thus the spectral excursion of noise peaks are reduced, and hence the musical noise lowered. The algorithm blockdiagram is shown in the Figure 1 [2].

3.2 Selection of value of α and β

Various combination 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 larger, the spectral floor is high and very little if any, musical

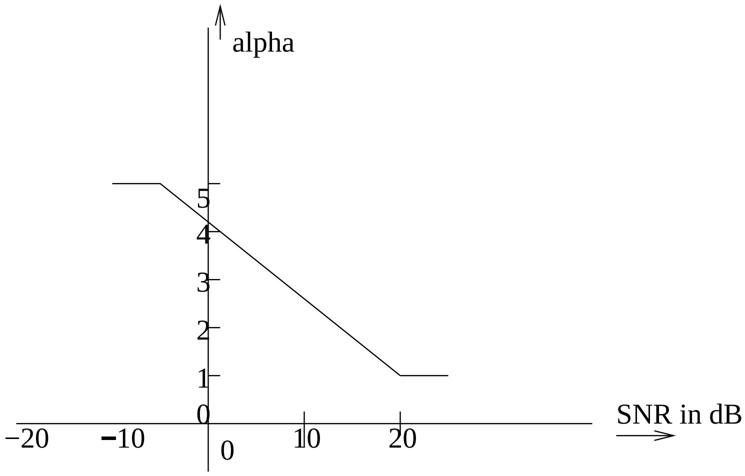

Figure 2 : Substraction factor α vs the SNR (Adapted from [2])

noise is audible. When β is small, the broad band noise is greatly reduced, but the musical noise becomes quite annoying. For a fixed value of β, increasing the value of α reduces both the broadband noise and musical noise. However if α is too large, the spectral distortion caused by the substraction in equation(8) becomes excessive and the speech intelligibility may suffer.

In order to reduce the speech distortion caused by large values of α, its value is adapted from frame to frame within the same sentence. The basic idea is to take into account that the substraction process must depend on the SNR of the frame in order to apply less substraction with high SNRs and vice-versa. This prevents unnecessary supression of speech signal when actual noise level is lower than our estimate. The reason for allowing α varying within the sentence is that the segmental SNR varies from frame to frame in proportional to the signal power. Value of α should vary within a sentence according to Figure 2 [2] with α=1 for SNR>20 dB and preventing α to increase further for SNR<-5 dB.

The slope of the line is determined by specifying the value of α at SNR=0 dB.Let p be the SNR. At each frame, actual value of α used in equation (9) is given by

$$\alpha = \alpha_0 - s.(p) \tag{10}$$

where α_0 is the value of α at SNR=0 dB,s is slope of line and SNR is signal to noise ratio in dB. It has been reported that using a variable substraction reduces speech distortion [2].

Thus we observed that there are several qualitative aspect of the processed speech that can be controlled. These are the level of the remaining broadband noise, the level of musical noise, and the amount of speech distortion. These 3 effects are controlled mainly by parametersα_0 and β.

3.3 Other Implementation Parameters

Apart from α and β discussed above, there are some other parameters in the substraction algorithm, which influence its performance [2]. These are
(1) The exponent of the spectrum of the input signal
(2) The frame size
(3) Window overlap
(1) Exponent of Spectrum

Instead of normal substraction of (9),the spectrum of input signal can be raised to some power γ before substraction, then (9) becomes

$$\text{Let } P(w) = |Y_w(w)|^\gamma - \alpha \left|\hat{D}_w(w)\right|^2$$

$$|\hat{S}_w(w)|^2 = [P^{2/\gamma}(w)] \quad if \;\; P^{2/\gamma}(w) > \beta \left|\hat{D}_w(w)\right|^2 \tag{11}$$

$$= \beta \left|\hat{D}_w(w)\right|^2 \text{ otherwise}$$

The value of γ being equal to 2 refers to the algorithm of equation (9). For fixed value of α, the substraction in (10) with the value of $\gamma < 2$ results in a greater change than for the case when $\gamma = 2$. Berrouti [2] concluded that the value $\gamma = 2$ is optmium for good quality of processed speech.

(2) Frame Size

If the frame size is large, musical noise [2] decreses. But if framesize is too large, then the speech signal can't be considered as a stationary signal within the frame and hence slurring effect occurs. It also has been reported that using frame size below 20ms results in roughness [2].

(3) Window Overlap

Speech is windowed such that in the absence of spectral modification, if synthesis speech segment are connected together the resulting overall system reduces to identity. The overlap is necessary to prevent discontinuities at frame boundries. The amount of overlap is taken to be 10% of the frame size [2].

3.4 Speech silence detection

The speech silence detector is an important part of the system for supression of additive noise in speech, because the quality of the detector determines the performance of the whole noise supression system [5]. Basic steps involved in the algorithm of different methods are

- Signal is divided into overlap frames.
- Some parameters of speech such as average magnitude, zero crossing are calculated.
- Thresholds are defined for making the decision between speech and silence.
- Decision is taken on some properties of speech, value of the parameter for the frame and threshold.

4. EVALUATION OF PERFORMANCE

The two main aspects of human perception of speech are speech intelligibility and quality. The quality of speech signal is a subjective measure which reflects the way the signal is perceived by listener. It can be expressed in terms of how pleasant the signal sounds or how much effort is required on behalf of listeners in order to understand the message. Intelligibility on the otherhand is an objective measures of the amount of information which can be extracted from by the listeners from the given signal, whether the signal is clean or noisy. A given signal may be of high quality but low intelligibility and vice-versa. Hence the 2 measures are independent of each other.

Traditionally, speech intelligibility is measured using Diagnostic Rhyme Test [5]. This subjective listening testing uses a set of isolated words to test consonant intelligibility.

It is also possible to measure intelligibility using sentence-level tests. Sentences taken from standard databases or nonsense sentences constructed from a specific syntactic structure are presented to the listeners who are asked to identify the words in each sentence [5].

Spectral substraction offers a computationally efficient approach for speech enhancementof speech degraded by additive background noise [1]. It improves the quality of the speech significantly by reducing background noise. Musical noise is also low when additional steps are taken. But when the aspect of intelligibility is considered any spectral substraction method based technique for speech enhancement in general doesn't improve the intelligibility and at times even lowers it. But in the context of using the method as preprocessor noise suppression technique in speech compression application, it not only improves the quality but also the speech intelligibility.

Boll [1] reported overall significant improvement in intelligibility and quality when spectral substraction technique was used as a preprocessor to an LPC speech analysis-synthesis vocoder, where they used DRT tests to evaluate the performance.

The intelligibility of the enhanced speech depends on the SNR. Many works reported that,the intelligibility decreases with decresing SNR, first gradually and then rapidly [5]. Coming to the modified spectral substraction method, Berouti [2] reported that at SNR= +5 dB, using the values α_0=3, β= 0.005 and γ=1, the intelligibility of the enhanced speech is same that of the unprocessed speech and degraded a little for lower SNR values. But it may be possible to fine tune the value of α_0 and β to get improvement in quality without loss in intelligibility, as they decide the the amount musical noise and background broadband noise remaining in the enhanced speech. The actual values are application specific, in some application, we may need more quality at the expense of slight loss in intelligibility and in some applications loss in intelligiblity may not be acceptable [2].

5. CONCLUSION

The various approaches applicable to single channel enhancement methods have been discussed briefly. We studied the basic method of sectral substraction and the reasons for its shortcomings of residual noise and musical noise have been discussed. Spectral substraction method has been a popular choice for removing additive stationary background noise mainly because of its simplicity in implementing the algorithm. Also it is computationally efficient requiring about the same computation as a high speed convolution [1]. Later a modified version of the spectral substraction method proposed by Berouti [2], which minimises the effect of musical noise and residual noise has been studied. In the modified method by substracting an overestimate of the noise power spectrum and preventing the resultant spectrum from going below a preset minimum level, both the musical and residual noise are minimised. The implementation aspects and the selection of parameters for minimising the residual and musical noise have been explained. Main limitation of the spectral substraction method lies in the fact that it degrades the intelligibility of the enhanced speech, especially at low SNR value. However this method is useful for pre-processing of noisy speech for digital speech processors used for speech compression, compression, recognition and authentication, since in this context it improves both quality and intelligibility.

The efficiency of present spectral substraction method depends on the speech-silence detection algorithm's accuracy. In the future versions of spectral substraction method, its performance should either be made independent of the accuracy of the speech-silence detection or the speech silence detection itself should be made further robust to improve the performance. Also, the technique should be able to suppress the non-stationary noise effectively. Further work should to be carried out to see that this method does not lower the intelligibility even at low SNR values.

ACKNOWLEDGEMENT

I would like to thank my guide, Prof. P.C. Pandey, for his guidence and motivation which helped me in completing seminar work.

REFERENCES

1. Boll S.F. , "Suppression of acoustic noise in speech using spectral substraction", *IEEE Trans. On Acoustics, Speech and Signal Processing,* vol.ASSP-27, pp.113-120, Apr. 1979.

2. M.Berouti, R.Schwartz and J.Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing,* pp.208-211, Apr. 1979.

3. Douglas O'Shaughnessy," Speech Communications, Human And Machine", Universities Press, 2001.

4. David L.Wang and Jae S.Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. On Acoustics, Speech, and Signal Processing*, vol.ASSP-30, no.4, pp.679-681, Aug.1982.

5. Gautam Moharir, "Spectral substraction method for speech enhancement", M.Tech Thesis, Department of electrical engineering, I.I.T Bombay, Mumbai, India, Jan 2002.