M.Tech. Credit Seminar Report
Electronic Systems Group
EE Dept, IIT Bombay
Submitted November 2002

SINGLE CHANNEL ENHANCEMENT OF NOISY SPEECH

Kotta Manohar (02307428)
Supervisor: Prof. Preeti Rao

Abstract

The problem of enhancement of speech degraded by additive background noise has received considerable attention over the past decade. The reasons being, its wide range of applications and limitations of the available methods. Among the available single channel enhancement techniques, the spectral subtraction technique has been one of the widely adopted methods for suppressing stationary additive background noise. The greatest asset of spectral subtraction lies in its simplicity since all that is required is an estimate of the mean noise power. The basic idea of spectral subtraction is to obtain an estimate of the speech spectral level by subtracting the noise estimation from noisy speech. The spectral subtraction 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 the available different approaches. It is followed by classification of single channel enhancement techniques and brief overview of each method is given. The basic spectral subtraction method and a modified version that minimizes the shortcomings of the basic method have been discussed in detail. The criteria used for evaluating the performance of a speech enhancement method have been explained and conclusions are drawn on the performance of spectral subtraction method. Finally, the limitations of the present spectral subtraction method have been identified and suggestions for further improvement of the method have been put forward.

I. INTRODUCTION

In speech communication, the speech signal is always accompanied by some noise. In most cases the background noise of the environment where the source of speech lies, is the main component of noise that adds to the speech signal. Though the obvious effect of this noise addition is to make the listening task difficult for a direct listener, there are many more far reaching negative effects when we process the degraded speech for some other applications. A related problem is processing degraded speech in preparation for coding by a bandwidth compression system. Hence speech enhancement not only involves processing speech signals for human listening but also for further processing prior to listening. Main objective of speech enhancement is to improve the perceptual aspects of speech such as overall quality, intelligibility, or degree of listener fatigue. In our work, we study one such enhancement technique to enhance the quality of speech in the presence of additive, broadband acoustic noise.

A. Various Types of Speech Degradation

Speech can be corrupted with noise at any stage before it reaches the end listener. The different ways in which the speech can be degraded can be broadly categorized as below.

a) *At the source of speech*

When the source itself is in a noisy environment, the background noise gets added to the speech signal. The background noise may be noise-like such as aircraft cockpit, other moving vehicle or environmental sounds; or it may be speech-like, comprised of competing speakers. Besides this, noise effect can also arise due to room reverberation where the source of speech is present. Here the noise involved is multiplicative noise, where the noise gets convolved with the speech signal.

b) *During the transmission of speech signal*

Speech is many a times transmitted over transmission channels like telephone lines to reach the listener at a distance. During this transmission, due to the non-idealities of channel behavior, additional noise is added to the speech signal. Noise can also be added either during the data conversion done before the transmission or during the speech reproduction at the listeners end.

c) *Noise at the listener end*

Sometimes though the speech source may be in a quiet environment, the end listener may be in a highly noisy environment. Here also the listening fatigue arises as the quality of speech is apparently lowered. Hence speech enhancement is needed in this case also.

B. Need for Speech Enhancement

The interfering noise generally degrades the quality and intelligibility of speech. While the term intelligibility refers to the recognisability of actual content of speech, quality refers to the aspect of the speech that determines the ease with which one can understand the speech. This degradation of speech by noise creates problems not only for just interpersonal communication but more serious problems in applications in which decision or control is made on the basis of speech signal. Hands free voice control is one such application.

Thus the main objective of speech enhancement is ultimately to improve one or more perceptual aspects of speech such as overall quality and intelligibility. The different contexts in which the enhancement techniques are used are as follows: -

1. Enhancement of speech degraded by additive noise
2. Making speech immune to degradation by processing speech before noise affects it.
3. Enhancement of speech degraded by reverberation.
4. Enhancement of speech keeping in view the needs of physiologically impaired persons.

We confine ourselves to the methods applicable to the first category.

C. Classification of Enhancement Techniques

Broadly the enhancement techniques can be classified as single channel and dual channel or multi channel enhancement techniques. Single channel enhancement techniques apply to situations in which only one acquisition channel is available. In dual channel enhancement techniques, a reference signal for the noise is available and hence adaptive noise cancellation technique can be applied. Multi channel enhancement techniques employ microphone arrays and take advantage of availability of multiple signal inputs to our system, to make possible the use of phase alignment to reject the undesired noise components.

Section II starts with introduction of single channel enhancement techniques and then gives an overview of few of the single channel enhancement techniques. Section III gives an overview of one popular method known as the basic spectral subtraction technique in the first sections and then a modified spectral subtraction is discussed in detail. Criteria by which performance of an enhancement method is evaluated have been discussed in Section IV, and in particular the performance of spectral subtraction method has been described. Section V summarizes the report, besides adding few conclusions drawn regarding the future scope for improvement in the domain of spectral subtraction techniques.

II. SINGLE CHANNEL ENHANCEMENT OF NOISY SPEECH

When only a single acquisition channel is available, single channel enhancement techniques are used. This may be imposed by the system used (as telephone based applications) or by the availability of the desired signal (as prerecorded applications). They are especially interesting due to the simplicity in microphone installation but the major constraint of single channel methods is that there is no reference signal for the noise available. Therefore the power spectral density of the noise has to be estimated based on the available noisy speech signal only and this is what makes it a challenging task.

In all single channel enhancement techniques, we assume the available speech signal model as

$$y(n) = s(n) + d(n)$$

Where *s(n)* represents the pure speech signal, which is assumed to be a stationary signal whenever processing is done on a short time basis, *d(n)* is the uncorrelated additive noise and *y(n)* represents the degraded speech signal.

There is broad set of applications for the single channel enhancement techniques. A few typical applications in which single channel enhancement techniques are used are listed below.

1. Communication between pilot cockpit and air traffic controller
2. Telephone services
3. Preprocessing of degraded speech before speech coding
4. Mobile telephony
5. Hands free voice control
6. Dictation and data entry
7. Speech compression

Every single channel enhancement technique exploits one or the other property of speech signal to separate the noise. Though there is no clear-cut classification for the available methods, most of them will come under one of the categories discussed below [1].

1. Methods based on periodicity of voiced speech
2. Methods based on a speech model
3. Methods based on short time spectral amplitude estimation

In the following sections of II-A, II-B and II-C we shall discuss briefly the different methods available under the classification outlined above.

A. Methods based on periodicity of voiced speech

Basic speech production model can be thought of as a linear system, representing the vocal tract, driven by an excitation function that is a periodic pulse train for voiced sounds and wideband noise for unvoiced sounds. Some enhancement methods capitalize on the observation that waveforms of voiced sounds are periodic with a period that corresponds to the fundamental frequency [1]. Again here different approaches are possible based on the same fundamental principle. They are briefly discussed here.

a) *Adaptive Comb Filtering*

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 the time 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 bands. Using this concept, the comb filtering technique approach is to pass the harmonics of speech but reject the frequency components between the harmonics using a filter. The filter should adapt globally and locally due to time varying nature of speech, i.e. variations of pitch from speaker to speaker and variation of pitch of same speaker over his speech. Hence accurate pitch information is needed. In applications where the background noise is a competing speaker, in short time spectrum the periodicity of speech exhibits itself as local spectral peaks some of which are due to the main speaker and some others of which are due to a competing speaker. We can separate the spectral peaks of the actual speaker and the competing speaker and then speech is regenerated.

b) *Adaptive Noise Canceling Techniques*

Besides the degraded signal, a reference signal that is uncorrelated with 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 subtracted from the degraded speech. But since we have only single acquisition channel, we have to generate the reference signal from the available degraded speech signal. In this technique, the reference input is generated by exploiting the periodicity of voiced speech.

In our assumed speech model

$$y(n) = s(n) + d(n)$$

If the speech signal is assumed periodic due to the periodicity of voiced speech, then the reference signal *r(n)* can be generated as

$$r(n) = y(n) - y(n-T)$$

If the periodicity is strictly observed, *r(n)* is uncorrelated with *s(n)* but is highly correlated with *d(n)* thus satisfying the conditions for adaptive noise canceling. But once again, we need accurate

estimate of the pitch to implement this; else good reference signal which is uncorrelated with the speech signal cannot be obtained. Though there are many conventional pitch extraction methods like linear prediction, autocorrelation and the cepstrum, they perform badly when extracting pitch of a degraded speech. So, new approaches have been proposed to extract pitch of degraded speech efficiently, one such approach is that has been particularly successful is the maximum likelihood pitch estimator proposed by Wise [1]. The methodology adopted in this approach to get the optimum pitch is to construct a bank of comb filters each tuned to a slightly different pitch period and choosing as the estimate the pitch corresponding to the comb filter for which the output energy is the largest.

In spite of many techniques being available for pitch estimation of noisy speech, due to time varying nature of speech even on a short time basis, the adaptive noise canceling technique is usually preferred only for dual channel enhancement techniques where we have access to a good reference signal easily.

B. Methods Based on a Speech Model

A simple but useful model for speech production consists of a time varying linear system representing the vocal tract, driven by an excitation function that is a periodic pulse train for voiced sounds and wide band noise for unvoiced sounds. The time varying linear system is akin to a digital filter whose coefficients change with time. However, over a short interval, the digital filter can be approximated as a linear time invariant system.

In these enhancement methods, parameters of the speech model are first estimated and later the speech is generated. The parameters of the model include source parameters (pitch information) and system parameters (vocal tract information). Speech can be generated by either one of the methods listed below.

1. A synthesis system based on the same speech model.
2. By designing a filter with estimated speech model parameters and then filtering the noisy speech.

Special methods are adopted to find source parameters (pitch estimation) of a noisy speech. The techniques to estimate the system parameters depend on the vocal tract model we adopt. The vocal tract can be modeled as an all-pole model or pole-zero model. After modeling in a chosen way, the parameters are estimated using one of the many available techniques like linear predictive analysis, or well-known parameter estimation rules like maximum likelihood (ML), minimum mean-square error (MMSE).

C. Methods Based On Short Time Spectral Amplitude Estimation

An aspect of human perception is that the short time spectral amplitude is more important than the phase for intelligibility and quality. This is exploited in these enhancement techniques, where only the spectral amplitude associated with the original signal is estimated, which is easier than to estimate both amplitude and phase. The techniques using this methodology can be broadly classified as below [1].

a) Speech Enhancement Based on Direct Estimation of Short-Time Spectral Amplitude

In this approach short time amplitude spectrum of speech is estimated in frequency domain, using spectrum of degraded speech and then short time segment of enhanced speech waveform in time domain is obtained by inverse transforming the amplitude estimate combined with the phase of degraded speech. Various enhancement techniques are available which are based on this concept, they differ primarily in how the short time energy spectrum of the original speech is estimated from the noisy speech. The whole class of spectral subtraction techniques comes under this category. This technique is discussed in detail in the next chapter.

b) Speech Enhancement Techniques Based on Weiner Filtering

The methodology here is to obtain a filter first from the noisy speech, which is then applied to the same degraded speech to get the enhanced speech. Here also only the spectral amplitude is enhanced retaining the phase of the noisy speech as the phase of the filtered speech. A frequency weighting for optimum filter is estimated from the noisy speech. The filter is then obtained in time domain or frequency domain to get the undegraded speech estimate. The frequency weighting for such a filter can be found based on the principles of Weiner filtering. If $Y_w(w), D_w(w)$, and $S_w(w)$ denote the

short-time spectra of the windowed time functions $y(n)$, $d(n)$ and $s(n)$, the estimate $\hat{S}_w(w)$ of $S_w(w)$ is given by

$$\hat{S}_w(w) = H(w)Y_w(w) \tag{1}$$

In the previous enhancement method based on direct estimation of short-time Spectral amplitude, we estimate short-time spectral magnitude explicitly through a process of spectral subtraction, whereas in this method, a Wiener filter that gives $\hat{S}_w(w)$, the estimate of $S_w(w)$ on applying to the degraded speech is found out first and then applied to the degraded speech.

An approach often used to find the frequency response of the Wiener filter is to approximate the noncausal wiener filter with an adaptive Wiener filter with frequency response

$$H(w) = \frac{E[|S_w(w)|^2]}{E[|S_w(w)|^2] + E[|D_w(w)|^2]} \tag{2}$$

Though we do not speak of any filter with frequency response *H(w)* in the previous method, the techniques in that method also can be put in the form represented by (1). Hence the basic difference between the two methods lies in the methodology adopted to obtain the frequency weighting *H(w)*. In this method, it is found by using the principles of Wiener filtering.

III. SPECTRAL SUBTRACTION METHOD

In the earlier chapter we have discussed the various methods available to deal with the problem of single channel enhancement of noisy speech. Among all these methods spectral subtraction method is the most popular choice when one has to eliminate the background stationary noise. The greatest asset of spectral subtraction lies in its simplicity, and the fact that all that is required is an estimate of the mean noise power and that the algorithm doesn't need any signal assumptions. At the same time the latter is its great weakness. Within the framework occasional negative estimates of the power spectrum can occur. To make the estimates consistent some artificial flooring is required, which yields a very characteristic musical noise, caused by the remaining isolated patches of energy in the time-frequency representation. But with suitable additional steps taken we can minimize the musical noise.

Spectral subtraction technique cannot only be used for enhancement of noisy speech for a direct listener, but also as a preprocessor noise reduction technique for digital voice processors used for speech compression, recognition and authentication. A typical example is spectral enhancement preprocessing for an LPC speech analysis-synthesis vocoder. In this section we will first discuss the basic spectral subtraction method and then a modified spectral subtraction method is studied in detail which has been suggested by Berouti [2]. The algorithm's strengths and limitations are identified.

A. Overview of the Basic Spectral Subtraction 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 sum of the signal power spectrum and noise power spectrum [1], by considering the speech signal model that we have already discussed in the earlier sections.

$$y(n) = s(n) + d(n)$$

Where *s(n)* represents the actual speech signal, *d(n)* is the 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 signals as processing is done on a short time basis. Above all, this method exploits the assumption that, for human perception the short time spectral amplitude is more important than the phase for intelligibility and quality. This assumption has been shown to be true, by many works, one such work is of Lim and Wang [6], wherein they observed that using the actual phase rather than the degraded speech phase does not improve the quality of the enhanced speech.

Considering the signal model

$$y(n) = s(n) + d(n)$$

Taking Fourier transforms, we get

$$Y(\omega) = S(\omega) + D(\omega)$$

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

$$y_w(n) = s_w(n) + d_w(n)$$

and its short-time Fourier transform is given by

$$|Y_w(\omega)|^2 = |S_w(\omega)|^2 + |D_w(\omega)|^2 + S_w(\omega).D_w^*(\omega) + S_w^*(\omega).D_w(\omega) \quad (3)$$

where $D_w^*(\omega), S_w^*(\omega)$ represent complex conjugates of $D_w(\omega)$ and $S_w(\omega)$. The function $|S_w(\omega)|^2$ is referred as the short-time energy spectrum of the speech. In the spectral subtraction method [1], the terms $|D_w(\omega)|^2, S_w(\omega).D_w^*(\omega)$ and $S_w^*(\omega).D_w(\omega)$ which cannot be obtained directly are approximated by $E\left[|D_w(\omega)|^2\right], E\left[S_w(\omega).D_w^*(\omega)\right]$ and $E\left[S_w^*(\omega).D_w(\omega)\right]$ where the $E[.]$ denotes the ensemble average. Since we have assumed the noise to be uncorrelated with the signal, $E\left[S_w(\omega).D_w^*(\omega)\right]$ and $E\left[S_w^*(\omega).D_w(\omega)\right]$ are zero and an estimate $|\hat{S}_w(\omega)|^2$ of $|S_w(\omega)|^2$, is suggested from (3) as

$$|\hat{S}_w(\omega)|^2 = |Y_w(\omega)|^2 - E\left[|D_w(\omega)|^2\right] \quad (4)$$

where $E\left[|D_w(\omega)|^2\right]$ is obtained either from the assumed known properties of *d(n)* or by an actual measurement from the background noise in the intervals where the speech is not present. Almost all speech enhancement methods based on spectral subtraction method follow the second option using the speech silence detection algorithms.

Since the estimate $|\hat{S}_w(\omega)|^2$ can become negative due to over-estimation of noise, it is made equal to zero in those cases.

Let $|\hat{D}_w(\omega)|^2 = E\left[|D_w(\omega)|^2\right]$ be the estimate of noise.

Hence complete power spectrum subtraction method is mathematically given by

$$\begin{aligned} |\hat{S}_w(\omega)|^2 &= |Y_w(\omega)|^2 - |\hat{D}_w(\omega)|^2 \qquad \text{if } |\hat{S}_w(\omega)|^2 > 0 \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned} \quad (5)$$

Going by our assumption that short time phase is relatively unimportant, we approximate $\mathrm{R}S_w(\omega)$, the phase of $S_w(\omega)$, by $\mathrm{R}Y_w(\omega)$, so that

$$\hat{S}_w(\omega) = |\hat{S}_w(\omega)|.\exp\left[\, j\mathrm{R}Y_w(\omega)\right] \quad (6)$$

and

$$\hat{s}_w(n) = F^{-1}\left\{\hat{S}_w(\omega)\right\} \quad (7)$$

The major problem of the basic method is that the algorithm may itself introduce a synthetic noise, called Musical Noise that arises due to the randomness of noise. Musical tones resulting from the subtraction of an average noise spectrum from non-stationary frame-level spectrum results in this. This sounds metallic and distracts the attention of the listener. To explain its existence one must note the fact that we subtract a smoothed estimate of noise spectrum whereas the short time power spectrum of actual white noise includes peaks and valleys [2]. Thus, after subtraction there remain peaks in the noise spectrum. Of those the narrower peaks which are large spectral excursions because

of the deep valleys that define them, are perceived as time varying notes which we refer to as musical noise. The wider ones are perceived as time varying broad band noise.

Besides the above problem of musical noise, though the noise is reduced, considerable broadband noise still remains in the processed speech. This broadband noise which remains in the enhanced speech after processing due to the inherent mismatch between the noise spectrum estimate and actual noise content at any instant is referred as "Residual Noise".

Added to these problems, sometimes speech can be identified as noise and suppressed since the characteristics of speech and noise are so similar that it is a challenging task distinguishing between them, leading to speech distortion.

B. A Modified Spectral Subtraction Method

As mentioned in the previous section, we need to modify the basic spectral subtraction method, to minimize the presence of residual noise and musical noise in the processed speech. Many modified spectral subtraction methods have been suggested and implemented over the past decade. Here we discuss one such method suggested by Berouti [2].

This method improves the broadband noise suppression better than basic spectral subtraction technique and also eliminates the musical noise. Besides it adapts to wide range of signal to noise ratios. There is no loss of speech intelligibility with enhancement.

a) Algorithm

Modifications made to the original spectral subtraction method are subtracting an overestimate of the noise power spectrum and preventing the resultant spectrum from going below a preset minimum level (spectral floor). These modifications lead to minimizing the perception of the narrow spectral peaks by decreasing the spectral excursions and thus lower the musical noise perception.

Thus the algorithm of basic method given by (5) is modified as

$$\begin{aligned} \left|\hat{S}_w(\omega)\right|^2 &= \left|Y_w(\omega)\right|^2 - \alpha\left|\hat{D}_w(\omega)\right|^2 \qquad \text{if } \left|\hat{S}_w(\omega)\right|^2 > \beta\left|\hat{D}_w(\omega)\right|^2 \\ &= \beta\left|\hat{D}_w(\omega)\right|^2 \qquad \text{otherwise} \end{aligned} \tag{8}$$

$$\text{with } \alpha \geq 1 \text{ and } 0 \leq \beta = 1$$

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

It can be seen from (8) that spectral noise peaks will be lower with $\alpha > 1$ than when $\alpha = 1$, i.e. the basic method approach. Also, with $\alpha > 1$ the subtraction can remove all of the broadband noise by eliminating most of the wide peaks. But deep valleys surrounding the narrow spectrum will remain in enhanced speech. This can be further reduced by filling-in the valleys. This objective is achieved by the keeping the spectral floor, $\beta\left|\hat{D}_w(\omega)\right|^2$. For $\beta > 0$, the valleys between the peaks are not as deep as for the case $\beta = 0$. Thus the spectral excursions of noise peaks are reduced, and hence the musical noise lowered. The algorithm block diagram [3] is shown in fig 1.

b) Implementation Aspects

An important implementation aspect is the speech silence detection. The speech silence detector is a part of the system, because the quality of the detector determines the performance of the whole enhancement system. Different algorithms are available for speech silence detection. Speech silence detection based on average magnitude and average zero crossing is one such algorithm. The speech silence detection is done on a frame to frame basis. Some principle causes of silence detection failure are weak fricatives or voiced fricatives that become unvoiced at the end of word, weak plosives at the end of word, nasal at the end of word [3]. Hence the efficiency of the speech silence detection algorithm applied decides the performance of the spectral subtraction enhancement technique.

Apart from α and β, there are some other parameters which influence its performance [2]. These are discussed next.

1. Exponent of Spectrum

Instead of normal subtraction of (8), the spectrum of input signal can be raised to some power γ before subtraction, then (8) becomes

$$Let\ P(\omega) = |Y_w(\omega)|^{\gamma} - \alpha |\hat{D}_w(\omega)|^{\gamma}$$

$$|\hat{S}_w(\omega)|^2 = \left[P^{2/\gamma}(\omega) \right] \quad \text{if } P^{2/\gamma}(\omega) > \beta |\hat{D}_w(\omega)|^2 \qquad (9)$$

$$= \beta |\hat{D}_w(\omega)|^2 \quad \text{otherwise}$$

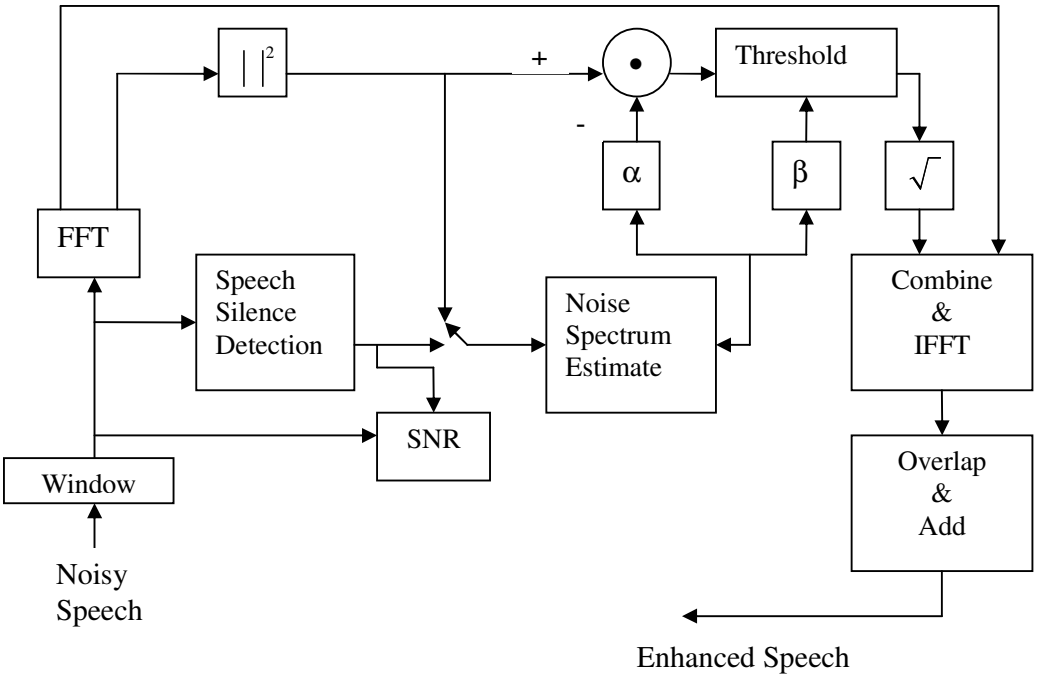

Fig. 1 Block diagram of modified spectral subtraction method ("from [3]")

The value of $\tilde{a}$ being equal to 2 refers to the algorithm of equation (8). For a fixed value of α, the subtraction in (9) with a value of $\tilde{a} < 2$ results in a greater change than for the case when $\tilde{a} = 2$. Berouti [2] concluded that value of $\tilde{a}=2$ is optimum for good output quality of processed speech.

2. Frame Size

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

3. Selection of α and β values

The selection of α and β values give rise to a trade-off between the amount of remaining residual broadband noise and the level of perceived musical noise. For β large, the spectral floor is high, and hence the musical noise will be reduced to a greater extent. While with β small, the residual broadband noise is greatly reduced but the musical noise becomes quite annoying [2]. On the other hand, for a fixed value of β, increasing the value of α reduces both the broadband residual noise and the musical noise. However, if α is too large the spectral distortion caused by the subtraction becomes excessive and the speech intelligibility may suffer.

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 subtraction process must depend on the SNR of the frame, in order to apply less subtraction with high SNRs and vice versa. This prevents unnecessary suppression 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. The α should vary within a sentence according to Fig. 2 [2], with α =1 for SNR > 20 and preventing α to increase further for SNR< -5.

The slope of the line is determined by specifying the value of α at SNR = 0 dB. At each frame actual value of α used in equation (8) is given by

$$\alpha = \alpha_0 - s \times (\text{SNR}) \tag{10}$$

Where α_0 is value of α at SNR = 0 dB, s is slope of line and SNR is signal to noise ratio in dB. It has been observed that it gives better results than keeping α constant [2].

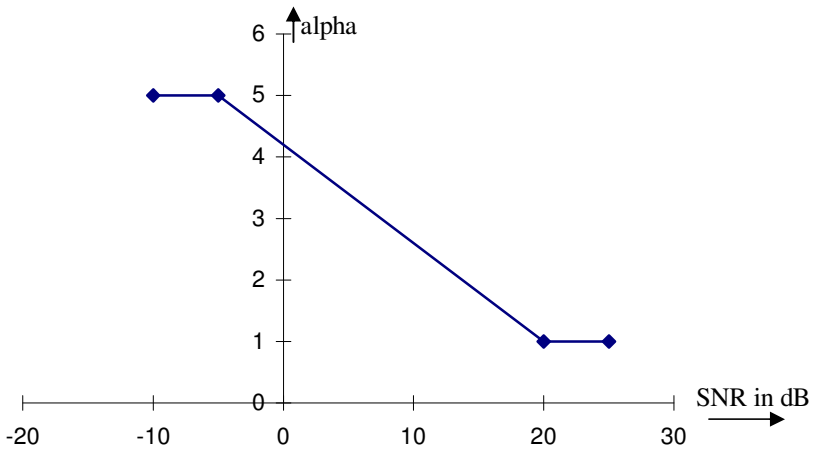

Fig. 2 Subtraction factor α Vs the SNR ("from [2]")

Thus we observe that there are several qualitative aspects of the speech that can be controlled, which are the level of the remaining residual broadband noise, the level of the musical noise and the amount of speech distortion. All these effects are controlled mainly by the parameters α_0 and β.

In the next section of IV, we shall discuss how the performance of a speech enhancement method is evaluated in general and also comment on the performance of spectral subtraction method in particular.

IV. EVALUATION OF PERFORMANCE

A. Introduction

With many speech enhancement methods being available today, a proper choice of method for a given application can be made only when we are able to evaluate performance of a speech enhancement method accurately. Since the human perceptual domain is not entirely well understood until now, we have to rely mainly on subjective tests such as a speech intelligibility or speech quality test. A careful subjective test can be tedious and time consuming, and generally requires processing a large amount of data. Hence some speech enhancement methods are evaluated on the basis of objective test such as S/N ratio when it is difficult to do a subjective test. Above all, the complicating factor in evaluating an enhancement method performance is that the objective of various methods available is generally an improvement in some aspects of the human perception such as an improvement in speech intelligibility or quality for a particular application under consideration.

Some methods which improve speech quality may reduce the speech intelligibility. Some methods which are used as preprocessor noise reduction techniques for digital voice processors can improve some aspects of speech only after the speech is processed, they may not be applicable for direct enhancement of noisy speech for a listener. A typical example is use of spectral subtraction enhancement preprocessing for a sinusoidal based codec [5]. It was reported that while for noisy speech enhancement, improving speech quality always accompanied a decrease in intelligibility, in the context of coding, significant combined improvements are obtained when the speech coder is

combined with a speech enhancement preprocessor. Thus performance evaluation of various available methods and their relative grading is a difficult task, but one can evaluate the performance of any method in the context of its application. In this chapter, we discuss how the main issues of human perception, the speech intelligibility and quality are measured and later comment on the performance of spectral subtraction method.

B. Measuring Speech Quality and Intelligibility

As mentioned in the earlier section, the two main aspects of human perception of speech are speech intelligibility and quality; an algorithm's performance is decided by these two parameters of the enhanced speech. The quality of speech signals is a subjective measure which reflects the way the signal is perceived by listeners [3]. It can be expressed in terms of how pleasant the signal sounds or how much effort is required on behalf of the listeners in order to understand the message. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted from 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 two measures are independent of each other. Both the quality and intelligibility must be evaluated on tests on human listeners as no closed form perceptually meaningful objective distortion measures are known.

Speech intelligibility is usually expressed as a percentage of words, sentences or phonemes correctly identified by a listener or group listeners when spoken by a talker or a group of talkers. It is an important measure of the effectiveness of a communication system in noisy environments. Listeners may have great difficulty in perceiving speech when it is heard in the presence of noise. This is because much of the acoustic information, which characterizes them, is often of low intensity and brief duration.

Traditionally, speech intelligibility is measured using the Diagnostic Rhyme Test (DRT) [3]. This subjective listening testing uses a set of isolated words to test consonants intelligibility. The test consists of 96 word pairs that differ by a single acoustic feature in an initial consonant. Word pairs are chosen to evaluate the six phonetic characteristics listed in Table I.

Characteristics	Description	Example
Voicing	Voiced-unvoiced	Veal-feel, dense-tense
Nasality	Nasal-Oral	Need-deed
Sustention	Sustained-interrupted	Vee-bee, sheet-cheat
Sibilation	Sibilated-unsibilated	Sing-thing
Graveness	Grave-acute	Weed-reed
Compactness	Compact-diffuse	Key-tea, show-so

TABLE I The DRT Characteristics ("from [3]")

The listener hears six words in a row, one from each category. In an answering sheet two options are given for each word. Then listener has to mark on the answering sheet which one of the two words he thinks is correct. Finally averaging the error rates from the answer sheets provides the results.

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 [3]. These semantically unpredicted sentences (SUS) are typically in the form of five different grammatical structures like

- Subject-Verb-Adverb
  The table walked through the blue truth.
- Subject-Verb-Object
  The strong way drank the day.
- Adverb-Verb-Object
  Never draw the house and the fact.
- Q-word-Verb-Subject-Object
  How does the day love the bright word?
- Subject-Verb-Object-Verb
  The plane closed the fish that lived.

But one has to note that for the same type and level of degradation, the intelligibility score can vary considerably depending on the test material, test procedure, training of subjects, etc [3]. When repeating the procedure with a listener, the test results may increase significantly due to learning effect which means the listener gets accustomed to the speech they hear and understand it better after every listening. Concentration problem, on the other hand, may degrade the results especially in sentence tests.

C. Performance of Spectral Subtraction Method

Spectral subtraction offers a computationally efficient approach for effective speech enhancement of speech degraded by additive background noise [4]. As we discussed in the earlier sections, it improves the quality of the speech significantly by reducing the background noise. Musical noise is also low when additional steps are taken. But when the aspect of intelligibility is considered, any spectral subtraction method based technique for speech enhancement in general does not improve the intelligibility and at times even lowers it. But in the context of using the method as a preprocessor noise suppression technique in speech compression applications, it not only improves the quality but also the speech intelligibility [1]. It has been supported by the results of many works where subjective tests were applied to evaluate the performance of the method. Boll [4] reported overall significant improvements in intelligibility and quality when spectral subtraction technique was used as a preprocessor to an LPC speech analysis-synthesis vocoder, where they used DRT tests to evaluate the performance. Besides subjective tests, one can also use short-time spectra, spectrograms or error measures to demonstrate the performance of the spectral subtraction method. Spectral subtraction enhancement preprocessing for a sinusoidal based codec [5], improved not only speech quality but also the speech intelligibility.

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

D. Future Scope for Improvement of the Spectral Subtraction Technique

Lots of research has been done in the area of noise reduction but even now it is a demanding task because of the following requirements of the system: -

1. Computational simplicity (i.e. low hardware component costs).
2. Very fast processing.
3. Quality and intelligibility of the enhanced speech.

Spectral Subtraction method has been a popular choice for removing additive stationary background noise mainly because of its simplicity in implementing the algorithm. Added it is computationally efficient requiring about the same computation as a high speed convolution [4]. But main limitation of the spectral subtraction lies in the fact that it degrades the intelligibility of enhanced speech, especially at low SNR values. Of course in the context of coding, it does improve intelligibility also.

The second shortcoming of the spectral subtraction method is that it itself introduces a synthetic noise called musical noise, which sounds metallic and distracts the listener. Though the modified spectral subtraction method reduces the musical noise, complete removal has not yet been realized. Added we note that in implementing the spectral subtraction technique, our efficiency also depends significantly on speech-silence detection algorithm, which itself needs further improvement in its accuracy. Thus there remains considerable further work to be done and room for improvement.

Significant improvement is still possible if one takes into account the possibility of nonstationarity of the noise signal, when formulating the algorithm. Then spectral subtraction method can also be used to suppress noise which is nonstationary in nature effectively.

VI. SUMMARY AND CONCLUSIONS

The need for enhancement of speech and the applications of speech enhancement methods has been discussed. The various approaches applicable to single channel enhancement methods have been discussed briefly. We studied in detail, the basic method of spectral subtraction method and the reasons for its main shortcomings of residual noise and musical noise have been elucidated. Later a modified version of the spectral subtraction method proposed by Berouti, which minimizes the effects of musical noise and residual noise has been studied. In the modified method by subtracting an overestimate of the noise power spectrum and preventing the resultant spectrum from going below a preset minimum level, the musical noise which arises due to randomness of the noise is minimized besides lowering the broadband residual noise. The implementation aspects and the selection of parameters for minimising the residual and musical noise have been explained.

The methods of evaluating the performance of a speech enhancement method have been studied and we discussed modified spectral subtraction method's performance in particular. In the work, the advantages and limitations of the spectral subtraction method when applied for noise suppression of additive background noise have been discussed. In summary, spectral subtraction improves quality of the speech signal significantly but when applied for normal noisy speech enhancement it will not improve the intelligibility in general and at times even may degrade the intelligibility at low SNR values. However the spectral subtraction method is a promising approach for pre-processing of noisy speech for digital speech processors used for speech compression, recognition and authentication, since in this context it not only improves the quality but also improves the intelligibility.

Though many modified versions of the spectral subtraction method have been reported over the time, there is still lot of scope for improvement in this technique. In the report earlier we commented that the present technique's efficiency depends on the speech-silence detection algorithm's accuracy. In the future versions of spectral subtraction, 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. Secondly the technique should be freed from the assumption of stationary noise and be able to suppress the non stationary noise effectively.

We conclude that the spectral subtraction method is one of the best choices available for single channel enhancement where the speech has been degraded by stationary additive background noise, especially as a pre-processor noise suppression system. Further work needs to be carried out to see that the method does not lower the intelligibility even at low SNR values and with no musical noise added. The basic speech signal modelling used in the current spectral subtraction method can be altered enabling it to suppress the non stationary noise also effectively.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to Prof. Preeti Rao for her constant guidance throughout the course of the work and many useful discussions which enabled me to know the subtleties of the subject in a proper way.

REFERENCES

1. J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. of the IEEE,* vol. 67, no.12, pp.1586-1604, Dec. 1979.

2. M. Berouti, R. Schwartz and J. Makhoul, “Enhancement of speech corrupted by acoustic noise”, *Proc. of the IEEE conference on Acoustics, Speech and Signal Processing,* pp.208-211, April 1979.
3. Gautam Moharir, “Spectral subtraction method for speech enhancement”, M.Tech thesis, Department of electrical engineering, I.I.T. Bombay, Mumbai, India, Jan 2002.
4. Boll S.F., “Suppression of acoustic noise in speech using spectral subtraction”, *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp.113-120, Apr 1979.
5. Gautam Moharir, Pushkar Patwardhan and Preeti Rao, “Spectral enhancement preprocessing for the HNM coding of noisy speech”, *Proc. of International Conference on Spoken Language Processing, Sep 2002.*
6. 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.