Survey of Noise Estimation Algorithms for Speech Enhancement Using Spectral Subtraction

Miss. Anuja Chougule,
Department of Electronic
Dr.JJMCOE,Jaysinghpur
Shivaji University,Kolhapur
anujachougule@gmail.com

Dr. Mrs. V. V. Patil,
Department of Electronics
Dr.JJMCOE,Jaysingpur
Shivaji University,Kolhapur
vvpatil2429@gmail.com

Abstract - Speech enhancement means speech improvement. Actually the speech enhancement is performed by using various techniques and different algorithms. Over the past several years there has been attention focused on the problem of enhancement of speech degraded by additive background noise. For many applications background suppression is required. The spectral - subtractive algorithm is one of the first algorithm proposed for additive background noise and it has gone through many modifications with time. For spectral subtraction method noise estimation is important for that there are various noise estimation algorithms. All these noise estimation algorithms are important for removing background noise.

Keywords: *Speech* Enhancement, *Spectral Subtraction, Noise Estimation*.

*****

I. INTRODUCTION

Broadband noise presented in speech signal can effect the quality of the signal, reduce intelligibility, and increase listener fatigue. Since in practice many types of noise is presented in recording speech, the problem of noise reduction is essential in the world of telecommunications.For noise reduction there are many algorithms are present.Noise reduction algorithms in general, attempt to improve the performance of communication systems when their input or output signals are corrupted by noise. The main objective of speech enhancement is to improve one or more perceptual aspects of speech, such as the speech quality or intelligibility. [1]

It is usually difficult to reduce noise without distorting speech The complexity and ease of implementation of any proposed system is another important criterion especially since the majority of the speech enhancement and noise reduction algorithms find applications in real-time portable systems like cellular phones, hearing aids, hands free kits etc.

There are many types of noise reduction algorithms have been developed mostly based on transform domain techniques, adaptive filtering, and model-based methods. Amongst the speech enhancement techniques, DFT-based transforms domain techniques have been used in the form of spectral subtraction

Even though the algorithm has very low computational complexity, it can reduce the background noise effectively. To reduce the influence of the background noise and increase the definition of the speech, the algorithm based on the modified spectral subtraction is introduced. Speech signals from the uncontrolled environment may contain degradation components along with required speech components. The degradation components include background noise, speech from other speakers etc.

SPEECH ENHANCEMENT

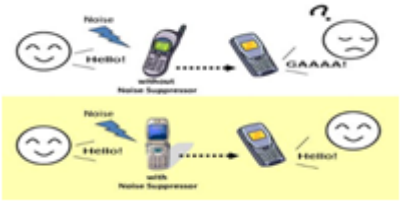


Figure 1:Introduction of speech enhancement.

Speech enhancement aims to improve speech quality by using various algorithms. It may sound simple, but what is meant by the word *quality*? Itcan be at leastclarity and intelligibility, pleasantness, or compatibility with some other method in speech processing.[2]

Intelligibility and pleasantness are difficult to measure by any mathematical algorithm. Usually listening tests are employed. However, since arranging listening tests may be expensive, it has been widely studied how to predict the results of listening tests. No single philosopher's stone or minimization criterion has been discovered so far and hardly ever will. The central methods for enhancing speech are the removal of background noise, *echo suppression* and the process of artificially bringing certain frequencies.

Another thing to remember is that quiet natural background noise sounds more comfortable than more quiet unnatural twisted noise. If the speech signal is not intended to be listened by humans, but driven for instance to a speech recognizer, then the comfortness is not the issue. It is crucial then to keep the background noise low. into the speech signal. We shall focus on the removal of background noise after briefly discussing what the other methods are all about.

II. SPECTRAL SUBTRACTION METHOD

Spectral Subtraction is a single channel speech enhancement technique. Single channel enhancement

 ISSN: 2321-8169
4156 - 4160

techniques apply to situations in which only one acquisition channel is available. These methods are interesting due to the simplicity [2]

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.

Basic Principles**:**

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

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

Where x(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.

Spectral subtraction is based on the principle that one can obtain an estimate of the clean signal spectrum by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The noise spectrum can be estimated, and updated, during the periods when the signal is absent or when only noise is present..

The noise corrupted input speech signal which is composed of the clean speech signal x (n) and the additive noise signal d (n) is shown in eq.above. The above eq. can be given in Fourier domain as shown:

$$Y[w] = X[w] + D[w]$$

Y[w] can be expressed in terms of Magnitude and phase as

$$Y[w] = Y\,|\,(w)|\; e\; j\; Ø\; y$$

Here |Y (w)| is the magnitude spectrum and Ø is the phase spectra of the corrupted noisy speech signal. Noise spectrum in terms of magnitude and phase spectra as

$$D[w] = |\,D[w]\,|\; e\; j\; Ø\; y$$

We can estimate the clean speech signal simply by subtracting noise spectrum from noisy speech spectrum, in equation form

$$X(w) = [|Y(w)| - |D(w)|]\; e\; j\; Ø\; y$$

The magnitude of noise spectrum |D (w)| is unknown but can be replaced by its average value computed during non speech activity i.e. during speech pauses.[3]

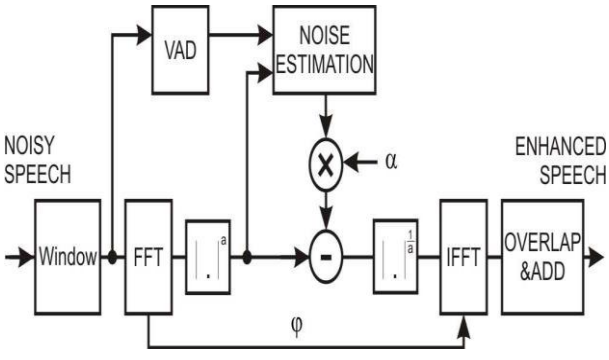


Figure 2 : Block diagram of spectral subtraction method.

Fig.2 shows a block diagram of the spectral subtraction method. The extent of the subtraction can be varied by applying a scaling factor a. The values of scaling factor a higher than 1 result in high SNR level of denoised signal, but too high values may cause distortion in perceived speech quality.

After subtraction, the spectral magnitude is not guaranteed to be positive. There are some possibilities to remove the negative components, for example half-wave rectification (setting the negative portions to zero), or full wave rectification (absolute value). Half-wave rectification is commonly used, but introduces musical tone artifacts in the processed signal. Full wave rectification avoids the creation of musical noise, but less effective at reducing noise.[3] After subtraction, *a* root of the X (w) is extracted to provide corresponding Fourier amplitude components. An inverse Fourier transform, using phase components directly from Fourier transform unit, and overlap add is then done to reconstruct the speech estimate in the time-domain.

A. Noise Estimation

A practical speech enhancement system consists of two major components, the estimation of noise power spectrum, and the estimation of speech. A critical component of any frequency domain enhancement algorithm is the estimation of the noise power spectrum. In single channel noise reduction/ speech enhancement systems, most algorithms require an estimation of average noise spectrum, and since a secondary channel is not available this estimation of the noise spectrum is usually performed during speech pauses. This requires a reliable speech/silence detector.

The speech/silence detection scheme can be a determining factor for the performance of the whole system. The speech/silence detection is necessary to determine frames of the noisy speech that contain noise only. Speech pauses or noise only frames are used for the noise estimate updating, making the estimation more accurate.

The decision about voice activity presence is the sensitive part of the whole spectral subtraction algorithm as the noise power estimation can be significantly degraded by the errors in voice activity detection. VAD accuracy dramatically affects the noise suppression level and amount of speech distortion that occurs.

B. Voice activity detection

Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected[4]. The main uses of VAD are in speech coding and speech recognition. It can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session: it can avoid unnecessary coding/transmission of silence packets in Voice over Internet Protocol applications, saving on computation and on network bandwidth. Independently from the choice of VAD algorithm, we must compromise between having voice detected as noise or noise detected as voice (between false positive and false negative). A VAD operating in a mobile phone must be able to detect speech in the presence of a range of very diverse types of acoustic background noise. In these difficult detection conditions it is often preferable that a VAD should fail-safe, indicating speech detected when the decision is in doubt, to lower the chance of losing speech segments. The biggest

 ISSN: 2321-8169
4156 - 4160

difficulty in the detection of speech in this environment is the very low signal-to-noise ratios (SNRs) that are encountered. VAD is an important enabling technology for a variety of speech-based applications. Therefore various VAD algorithms have been developed that provide varying features and compromises between latency, sensitivity, accuracy and computational cost. Some VAD algorithms also provide further analysis, for example whether the speech is voiced, unvoiced or sustained. Voice activity detection is usually language independent [2].The typical design of a VAD algorithm is as follows:

1. There may first be a noise reduction stage, e.g. via *spectral subtraction*.
2. Then some features or quantities are calculated from a section of the input signal.
3. A classification rule is applied to classify the section as speech or non-speech – often this classification rule finds when a value exceeds a threshold.

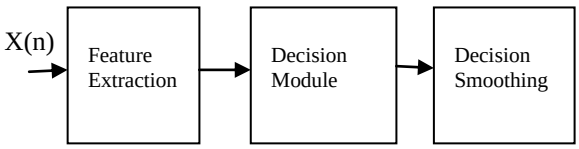


Figure 3: Block diagram of VAD

Applications-

1) VAD is an integral part of different speech communication systems such as audio conferencing, echo cancellation, speech recognition, speech encoding, and hands-free telephony.
2) In the field of multimedia applications, VAD allows simultaneous voice and data applications.
3) Similarly, in Universal Mobile Telecommunications Systems (UMTS), it controls and reduces the average bit rate and enhances overall coding quality of speech.
4) In cellular radio systems (for instance GSM and CDMA systems) based on Discontinuous Transmission (DTX) mode, VAD is essential for enhancing system capacity by reducing co-channel interference and power consumption in portable digital devices [5]

For a wide range of applications such as digital mobile radio, Digital Simultaneous Voice and Data (DSVD) or speech storage, it is desirable to provide a discontinuous transmission of speech-coding parameters. Advantages can include lower average power consumption in mobile handsets, higher average bit rate for simultaneous services like data transmission, or a higher capacity on storage chips. However, the improvement depends mainly on the percentage of pauses during speech and the reliability of the VAD used to detect these intervals. On the one hand, it is advantageous to have a low percentage of speech activity. On the other hand clipping, that is the loss of milliseconds of active speech, should be minimized to preserve quality.

III. LIMITATIONS OF SPECTRAL SUBTRACTION

A. *Impossibility to use for non-stationary noise.*

Noise spectrum estimate is obtained from the speech non-native regions of noisy speech. This assumption is valid for the case of stationary noise in which the noise spectrum does not vary much over time. Traditional VADs track the noise only frames of the noisy speech to update the noise estimate. But the update of noise estimate in those methods is limited to speech absent frames. This is not enough for the case of non-stationary noise in which the power spectrum of noise varies even during speech activity.[4]

B. *Dependence* on VAD accuracy.

Spectral subtraction performance is limited by the accuracy of noise estimation, which additionally is limited by the performance of speech/pause detectors. Performance of whole spectral subtraction noise reduction algorithm as well as VAD performance degrades significantly at lower SNR.

C. *Musical noise.*

Although spectral subtraction method provide an improvement in terms of noise attenuation, it often produce a new randomly fluctuating type of noise, referred to as musical noise due to their narrow band spectrum and presence of tone-like characteristics. This phenomenon can be explained by noise estimation errors leading to false peaks in the processed spectrum.

When the enhanced signal is reconstructed in the time-domain, these peaks result in short sinusoids whose frequencies vary from frame to frame. Musical noise although very different from the original noise, can sometimes be very disturbing. A poorly designed spectral subtraction, which caused musical noise, can sometime results in the signal that has lower perceived quality and lower information content, than the original noisy signal. Most of the research at the present time is focused in ways to combat the problem of musical noise.

IV. NOISE ESTIMATION ALGORITHMS

There are various noise estimation algorithms. The noise estimation algorithms are used to obtain clear speech signal from noisy speech signal.

A. *Estimating the noise spectrum based on tracking the minimum of the noisy speech*

Martin proposed a method for estimating the noise spectrum based on tracking the minimum of the noisy speech over a finite window. As the minimum is typically smaller than the mean, unbiased estimates of noise spectrum

were computed by introducing a bias factor based on the statistics of the minimum estimates. The main drawback of this method is that it takes slightly more than the duration of the minimum-search window to update the noise spectrum when the noise floor increases abruptly. [6]

B. Minima controlled recursive algorithm (MCRA)

A proposed a minima controlled recursive algorithm (MCRA) which updates the noise estimate by tracking the noise-only regions of the noisy speech spectrum. These regions are found by comparing the ratio of the noisy speech to the local minimum against a threshold. The noise estimate, however, lags by at most twice that window length when the noise spectrum increases abruptly. In the improved MCRA approach , a different method was used to track the noise-only regions of the spectrum based on the estimated speech-presence probability. This probability, however, is also controlled by the minima, and therefore the algorithm incurs roughly the same delay as the MCRA algorithm for increasing noise levels.[7]

C. Noise estimate by continuously tracking the noisy speech in each frequency bin.

In this algorithm the noise estimate is updated by continuously tracking the minimum of the noisy speech in each frequency bin. As such, it is computationally more efficient than the method in [6]. However, it fails to differentiate between an increase in noise floor and an increase in speech power. Hirsch and Ehrlicher updated the noise estimate by comparing the noisy speech power spectrum to the past noise estimate. Their method is also simple to implement, however it fails to update the noise estimate when the noise floor increases abruptly and stays at that level. [9]

D. Narrow-band spectral analysis

In this analysis there is a combination of the above techniques with narrow-band spectral analysis which allowed estimation of the noise levels in the valleys between harmonics of voiced speech segments. Longer time windows were required to achieve the required spectral resolution. Although their approach refines the spectral resolution of the noise level, it does not adapt faster to increasing noise levels. Lastly, in (Stahl et al., 2000) a quantile-based noise-estimation algorithm was proposed which estimates the noise spectrum based on the qth quantile of the noisy speech power spectrum. This method might fail to estimate the noise floor correctly if the noisy speech contains highly-varying noise.

E. Noise estimation algorithms for highly non-stationary noise environments.

The smoothed power spectrum of noisy speech is computed using the following first-order recursive equation:

$$P(\alpha,k) = gP(\alpha-1,k)+(1-g)|Y(\alpha,k)|2$$

Where $P(\alpha,k)$ is the smoothed power spectrum, α is the frame index, k is the frequency index, $Y|(\alpha,k)|^2$is the short-time power spectrum of noisy speech [10]
The proposed algorithm is summarized in the following steps.

1) Classification of Speech Present and Speech absent Frames**:**

In any speech sentence there are pauses between words which do not contain any speech; those frames will contain only background noise. The noise estimate can be updated by tracking those noise only frames .

To identify those frames , a simple procedure is used which calculates the ratio of noisy speech power spectrum to the noise power spectrum at 3 different frequency bands in each frame correspond to the frequency bins of 1 KHz, 3KHz and the sampling frequency respectively. If all the three ratios are smaller than the threshold that frame is concluded as a noise only frame, otherwise , if any one or all the ratios are greater than threshold that frame is considered as speech present frame. [11]

The noise estimate is updated in speech absent frames with a constant smoothing factor. In speech present frames the noise is updated by tracking the local minimum of noisy speech and the deciding speech presence in each frequency bin separately using the ration of noisy speech power to its local minimum.

2) Minimum of Noisy Speech:

Various methods were proposed for tracking the minimum of the noisy speech power spectrum over a fixed search window length. These methods were sensitive to outliers and also the noise update was dependent on the length of the minimum-search window. A different non-linear rule is used in our method for tracking the minimum of the noisy speech by continuously averaging past spectral values. In this algorithm if the value of the noisy speech spectrum in the present frequency bin is greater than the minimum value of previous frequency bin then the minimum value is updated, else the previous value is maintained as it is.[13]

3) Detection of Speech-Presence Frames:

The approach taken to determine speech presence in each frequency bin is similar to the method used in [4]. Let the ratio of noisy speech power spectrum and its local minimum be defined as

$$S_r(\alpha, k) = P(\alpha, k)/P_{min}(\alpha, k)$$

This ratio is compared with a frequency dependent threshold, and if the ratio is found to be greater than the threshold, it is taken as a speech-present frequency bin else it is taken as a speech-absent frequency bin. This is based on the principle that the power spectrum of noisy speech will be nearly equal to its local minimum when speech is absent. Hence the smaller the ratio is in, the higher the probability that it will be a noise only region and vice versa.

 ISSN: 2321-8169
4156 - 4160

4) Frequency-Dependent Smoothing Constants:

Using the above speech-presence probability estimate, we compute the time–frequency dependent smoothing factor as follows:

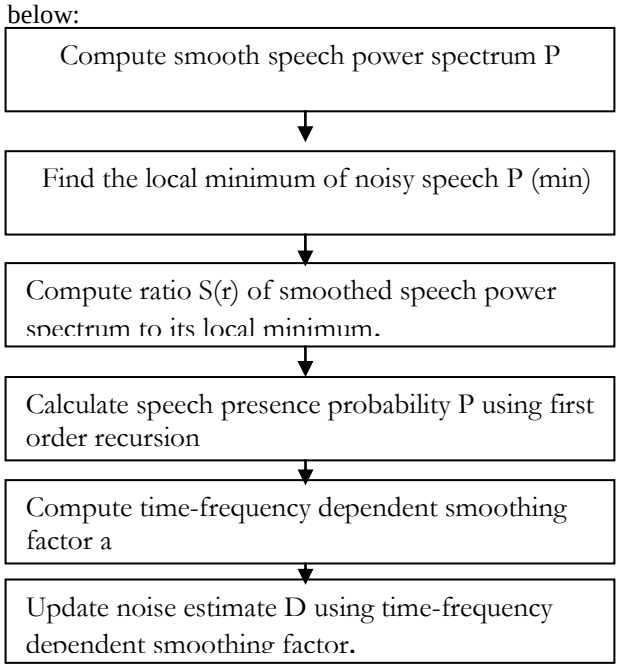
$$A(\alpha,k)=d+(1-d)P(\alpha,k)$$

5) Updating Noise Spectrum Estimate:

Finally, after computing the frequency-dependent smoothing factor a(α,k), the noise spectrum estimate is updated as

$$D(\alpha,k)=a(\alpha,k)D(\alpha-1,k)+(1-a(\alpha,k))|Y(\alpha,k)|^2$$

Where D(α,k) is the estimate of the noise power spectrum .Hence, the overall algorithm can be summarized as follows. After classifying the frequency bins into speech present/absent, we update the speech- presence probability and then use this probability to update the time– frequency dependent smoothing factor. Finally the noise spectrum estimate is updated using the time–frequency dependent smoothing factor. This estimated noise is then subtracted from the input noisy speech signal to get an estimate of clean speech. The flow diagram of above steps is given below:



V. DISCUSSION

Various spectral subtraction algorithms proposed for speech enhancement were described in above sections. These algorithms are computationally simple to implement as they involve a forward and an inverse Fourier transform. The simple subtraction processing comes at a price. The subtraction of the noise spectra from the noisy spectrum introduces a distortion in the signal known as Musical noise. We presented different techniques that mitigated the Musical noise distortion. Different variations of spectral subtraction were developed over the years.

The most common variation involved the use of an over subtraction factor that controlled to some amount of speech spectral distortion caused by subtraction process. Use of spectral floor parameter prevents the resultant spectral components from going below a preset minimum value. The spectral floor value controlled the amount of remaining residual noise and the amount of musical noise.

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise” in Processing of international Conference of Acoustic, Speech and Signal Processing,1979, pp. 208-211.

[2] V. Prasad, R. Sangwan et al., “Comparison of voice activity detection algorithms for VoIP”, proc. of the Seventh International Symposium on Computers and Communications, Taormina, Italy, 2002, pp. 530-532.

[3] K.Sakhnov, E.Verteletskaya, B. Šimák, “Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications”. In World Congress of Engeneering 2009 Proceedings. Hong Kong:, 2009, pp.801-806.

[4] Faneuff J.J, Brown D. R. “Noise Reduction and Increased VAD Accuracy Using Spectral Subtraction”, Processing of the Global Signal Processing Exposition and International Signal Processing Conference (ISPC’ 03). Dallas, Texas. April 2003.

[5] Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512

[6] Cohen, I., 2002. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. 9 (1), 12–15.

[7] Ramírez, J.; Górriz, J.M; Segura, J.C.; Puntonet, C.G; Rubio, A. (2006a). Speech/Non-speech

[8] Discrimination based on Contextual Information Integrated Bispectrum LRT, IEEE Signal Processing Letters, vol. 13, No. 8, pp. 497-500.

[9] Doblinger, G., 1995. Computationally efficient speech enhancement by spectral minima tracking in subbands. Proc. Eurospeech 2, 1513–1516.

[10] Górriz, J.M.; Ramírez, J.; Puntonet, C.G.; Segura, J.C. (2006a). Generalized LRT-based voice activity detector, IEEE Signal Processing Letters, Vol. 13, No. 10, pp. 636-639.

[11] Sohn, J.; Kim, N.S.; Sung, W. (1999). A statistical model-based voice activity detection, IEEE

[12] Signal Processing Letters, vol. 16, no. 1, pp. 1–3.

[13] Martin, R., 1994. Spectral subtraction based on minimum statistics. Proc. Eur. Signal Process., 1182–1185.

[14] Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11 (5), 466–475.

[15] T. Esch, P. Vary, ”Efficient musical noise suppression for speech enhancement system” IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 4409 - 4412.