www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 10 Oct 2015, Page No. 14873-14878

Speech Enhancement using Spectral Subtraction-type Algorithms: A survey on Comparison

Pinki[1], Sahil Gupta[2]
[1]M.Tech Scholar, Geeta Institute of Management and Technology, Kurukshetra
Pinkirana58@gmail.com
[2]Astt.Professor, Geeta Institute of Management and Technology, Kurukshetra
sahil.btech.ece@gmail.com

Abstract: ***The term "Speech Enhancement" refereed as to improve quality or intelligibility of speech signal. Speech signal is often degraded by additive background noise like babble noise, train noise, restaurant noise etc. In such noisy environment listening task is very difficult at the end user. Many times speech enhancement is used for pre processing of speech for computer speech recognition system. In this method, the noise spectrum is estimated during speech pauses, and is subtracted from the noisy speech spectrum to estimate the clean speech. This is also achieved by multiplying the noisy speech spectrum with a gain function and later combining it with the phase of the noisy speech. The drawback of this method is the presence of processing distortions, called remnant noise. A number of variations of the method have been developed over the past years to address the drawback. These variants form a family of spectral subtractive-type algorithms. The aim of this paper is to provide a comparison and simulation study of the different forms of subtraction-type algorithms viz. basic spectral subtraction, Modified Spectral Subtraction , multi-band spectral subtraction, iterative spectral subtraction, and spectral subtraction based on perceptual properties.***

Keywords: Speech enhancement, spectral subtraction, quality measure, Noise estimation, Spectral subtractive-type algorithms, Remnant noise, Objective evaluation.

1. Introduction

Speech communication is the exchange of information *via* speech either between humans or between human to machine in the various fields' for instance automatic speech recognition and speaker identification. In many Situations, speech signals are degraded by the ambient noises that limit their effectiveness of communication. Therefore enhancement of speech is normally required to reduce annoyance due to noise. The main purpose of speech enhancement is to decrease the distortion of the desired speech signal and to improve one or more perceptual aspects of speech, such as the quality and/or intelligibility. These two measures are not necessarily correlated. Therefore, an increase in speech quality does not necessarily lead to an improvement in intelligibility. Speech enhancement techniques can be classified into, single channel, dual channel or multi-channel enhancement. Although the performance of multi-channel speech enhancement is better than that of single channel enhancement, the single channel speech enhancement is still a significant field of research interest because of its simple implementation and ease of computation. In single channel applications, only a single microphone is available and the characterization of noise statistics is extracted during the periods of pauses, which requires a stationary assumption of the background noise. The estimation of the spectral amplitude of the noise data is easier than estimation of both the amplitude and phase. It is revealed that the short-time spectral amplitude (STSA) is more important than the phase information for the quality and intelligibility of speech. Based on the STSA estimation, the single channel enhancement technique can be divided into two classes. The first class attempts to estimate the short-time spectral magnitude of the speech by subtracting a noise

estimate. The noise is estimated during speech pauses of the noisy speech. The second class applies a spectral subtraction filter (SSF) to the noisy speech, so that the spectral amplitude of enhanced speech can be obtained. The design principle is to select appropriate parameters of the filter to minimize the difference between the enhanced speech and the clean speech.

1.1 SPECTRAL SUBTRACTION METHOD

1.1.1 Algorithm and Implementation:

Many different algorithms have been proposed for speech enhancement: the one that we will use is known as spectral subtraction. This technique operates in the frequency domain and makes the assumption that the spectrum of the input signal can be expressed as the sum of the speech spectrum and the noise spectrum. The procedure is illustrated in the diagram below and contains two tricky parts:

- estimating the spectrum of the background noise
- subtracting the noise spectrum from the speech

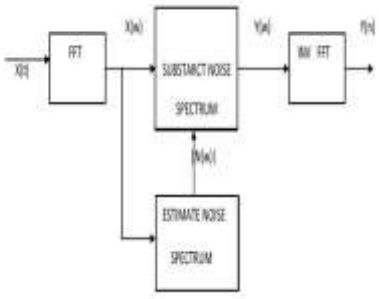

Fig 1: Block diagram of noise subtraction in spectral domain

1.1.2 Overlap-add processing

To perform frequency-domain processing, it is necessary to split the continuous time domain signal up into overlapping chunks called frames. After processing, the frames are then reassembled to create a continuous output signal. To avoid spectral effects, we multiply the frame by a window function before performing the FFT and again after performing the inverse-FFT.

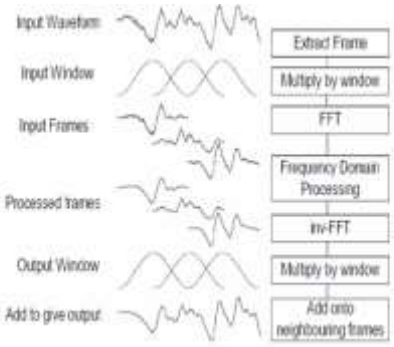

Fig 2: General steps of overlap-and-add processing on streaming data

2. Principle of Spectral Subtraction Method

Consider a noisy signal which consists of the clean speech degraded by statistically independent additive noise

$$y[n] = s[n] + d[n] \quad (1)$$

where y[n], s[n] and d[n] are the sampled noisy speech, clean speech, and additive noise, respectively. It is assumed that additive noise is zero mean and uncorrelated with the clean speech. Because the speech signal is non-stationary and time variant, the noisy speech signal is often processed on a frame-by-frame. Their representation in the short-time Fourier transform (STFT) domain is given by

$$Y(\omega, k) = S(\omega, k) + D(\omega, k) \quad (2)$$

Where k is a frame number. Throughout this paper, it is assumed that the speech signal is segmented into frames, hence for simplicity, we drop k.

Since the speech is assumed to be uncorrelated with the background noise, the short-term power spectrum of y[n]has no cross-terms. Hence,

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \quad (3)$$

The speech can be estimated by subtracting a noise estimate from the received signal.

DOI: 10.18535/ijecs/v4i10.43

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - |\hat{D}(\omega)|^2 \quad (4)$$

The estimation of the noise spectrum $|\hat{D}(\omega)|^2$ is obtained by averaging recent speech pauses frames:

$$|\hat{D}(\omega)|^2 = \frac{1}{M}\sum_{j=0}^{M-1}|Y_{SP_j}(\omega)|^2 \quad (5)$$

where M is the number of consecutive frames of speech pauses (SP). If the background noise is stationary, converges to the optimal noise power spectrum estimate as a longer average is taken.

The spectral subtraction can also be looked at as a filter, by manipulating (4) such that it can be expressed as the product of the noisy speech spectrum and the spectral subtraction filter (SSF) as:

$$|\hat{S}(\omega)|^2 = \left(1 - \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}\right)|Y(\omega)|^2 = H^2(\omega)|Y(\omega)|^2 \quad (6)$$

where H(ω) is the gain function and known spectral subtraction filter (SSF). The H(ω) is a zero phase filter, with its magnitude response in the range of $0 \le H(\omega) \le 1$.

$$H(\omega) = \left\{\max\left(0, 1 - \frac{|\hat{D}(\omega)|^2}{|Y(\omega)|^2}\right)\right\}^{1/2} \quad (7)$$

To reconstruct the resulting signal, the phase estimate of the speech is also needed. A common phase estimation method is to adopt the phase of the noisy signal as the phase of the estimated clean speech signal , based on the notion that short-term phase is relatively unimportant to human ears. Then, the speech signal in a frame is estimated as

$$\hat{S}(\omega) = |\hat{S}(\omega)|e^{j<Y(\omega)} = H(\omega)Y(\omega) \quad (8)$$

The estimated speech waveform is recovered in the time domain by inverse Fourier transforming S (ω) using an overlap and add approach.

The spectral subtraction method, although reducing the noise significantly, it has some severe drawbacks. From (4), it is clear that the effectiveness of spectral subtraction is heavily dependent on accurate noise estimation, which is a difficult task to achieve in most conditions. When the noise estimate is less than perfect, two major problems occur, remnant noise with musical structure and speech distortion.

2.1 Modified Spectral Subtraction

Modifications made to the original spectral subtraction method are subtracting an over estimate of the noise power spectrum and preventing the resultant spectrum from going below a preset minimum level (spectral floor).This modifications lead to minimizing the perception of the narrow spectral peaks by decreasing the spectral excursions and thus lower the musical noise effect. Berouti [4] has taken a different approach that does not require access to future information. This Method consists of subtracting an overestimate of the noise power spectrum and presenting the resultant spectral components from going below a preset minimum spectral floor value.

In this algorithm, two additional parameters are introduced in the spectral subtraction method over-subtraction factor, and noise spectral floor to reduce the remnant noise. The algorithm is given as

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\hat{D}(\omega)|^2, & \text{if } |Y(\omega)|^2 > (\alpha+\beta)|\hat{D}(\omega)|^2 \\ \beta|\hat{D}(\omega)|^2 & \text{else} \end{cases} \quad (9)$$

with $\alpha \ge 1$ and $0 \le \beta << 1$.

The over-subtraction factor controls the amount of noise power spectrum subtracted from the noisy speech power spectrum in each frame and spectral floor parameter prevent the resultant spectrum from going below a preset minimum level rather

than setting to zero (spectral floor). The over-subtraction factor depends on a-posteriori segmental SNR (SSNR). The over-subtraction factor can be calculated as

$$\alpha = 4 - \frac{3}{20}\,\mathrm{SSNR},\ \text{if } -5 \le \mathrm{SSNR} \le 20 \quad (10)$$

$$\mathrm{SSNR} = \left(\frac{\sum_{k=0}^{NF-1} |Y(\omega)|^2}{\sum_{k=0}^{NF-1} |\widehat{D}(\omega)|^2} \right) \quad (11)$$

Here NF is the number of frames in the signal.

2.2 Multi-band spectral subtraction

Real world noise is mostly colored and affects the speech signal differently over the entire spectrum. The plot of SSNR of non-overlapped uniformly spaced frequency bands {60Hz -1 kHz (Band 1), 1kHz -2kHz (Band 2), 2kHz-3 kHz (Band3), 3 kHz-4kHz (Band 4)} over frame number. This figure shows that the SSNR of the low frequency bands (Band 1) is significantly higher than the SSNR of higher frequency bands (Band 4) Therefore, the use of frequency dependent subtraction factor to account for different types of noise. The idea of non-linear spectral subtraction (NSS), basically extend this capability by making the over-subtraction factor frequency dependent and subtraction process is non-linear. Larger values are subtracted at frequencies with low SNR levels, and smaller values are subtracted at frequencies with high SNR levels. Certainly, this gives higher flexibility in compensating for errors in estimating the noise energy in different frequency bins. To take into account, a uniformly frequency spaced multi-band approach to spectral subtraction was presented in10. In this algorithm, the speech spectrum is divided into four uniformly spaced frequency bands, and spectral subtraction is performed independently in each band. The algorithm re-adjusts the over-subtraction factor in each band based on SSNR. So, the estimate of the clean speech magnitude spectrum in the ith Band is obtained by:

$$|\widehat{S}_i(\omega)|^2 = \begin{cases} |Y_i(\omega)|^2 - \alpha_i\,\delta_i |\widehat{D}_i(\omega)|^2, & \text{if } |\widehat{S}_i(\omega)|^2 > 0\, k_i < |\omega < k_{i+1} \\ \beta |Y_i(\omega)|^2 & \text{else} \end{cases} \quad (12)$$

Where k_i and k_{i+1} are the start and end frequency bins of the ith frequency band, α_i is the band specific over-subtraction factor of the ith Band, which is the function of SSNR of the ith frequency band. The SSNR of the ith frequency band can be calculated as

$$\mathrm{SSNR}_i(\omega) = \left(\frac{\sum_{\omega=k_i}^{k_{i+1}} |Y_i(\omega)|^2}{\sum_{\omega=k_i}^{k_{i+1}} |\widehat{D}_i(\omega)|^2} \right) \quad (13)$$

The band specific over-subtraction can be calculated, as

$$\alpha_i = \begin{cases} 5, & \text{if } \mathrm{SNR}_i \le -5 \\ 4 - \frac{3}{20}\,\mathrm{SSNR}_i, & \text{if } -5 \le \mathrm{SNR}_i \le 20 \\ 1, & \text{if } \mathrm{SNR}_i > 20 \end{cases} \quad (14)$$

The δ_i is an additional band subtraction factor that can be individually set for each frequency band to customize the noise removal process and provide an additional degree of control over the noise subtraction level in each band.

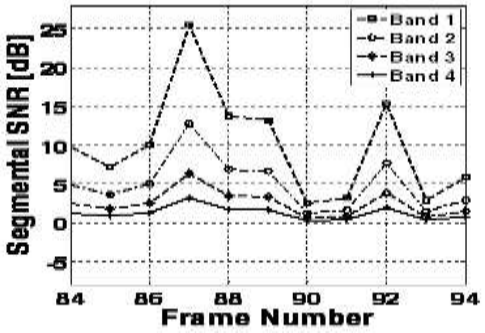

Fig 3 : The segmental SNR of bands

The values of δ_i is empirically calculated and set to

$$\delta_i = \begin{cases} 1, & f_i \le 1\,\mathrm{kHz} \\ 2.5, & 1\,\mathrm{kHz} < f_i \le \frac{f_s}{2} - 2\,\mathrm{kHz} \\ 1.5, & f_i > \frac{f_s}{2} - 2\,\mathrm{kHz} \end{cases} \quad (15)$$

Here f_i is the upper bound frequency of the ith Band and f_s is the sampling frequency. The motivation for using smaller values of δ_i for the low frequency

bands is to minimize speech distortion, since most of the speech energy is present in the lower frequencies. Both factors, alpha and δ_i can be adjusted for each band for different speech conditions to get better speech quality. As the real-world noise is highly random in nature, improvement in the MBSS algorithm for reduction of WGN is necessary. The MBSS algorithm is found to perform better than other subtractive-type algorithms.

2.3 Iterative spectral subtraction

An iterative spectral subtraction (ISS) algorithm is proposed in which is motivated from WF, to suppress the remnant noise. In this algorithm, the output of the enhanced speech is used as the input signal for the next iteration process. As after the spectral subtraction process, the type of the additive noise is transformed to the remnant noise and the output signal is used as the input signal of the next iteration process. The remnant noise is re-estimated and this new estimated noise, furthermore, is used to process the next spectral subtraction process. Therefore, an enhanced output speech signal can be obtained, and the iteration process goes on. If we regard the process of noise

3.2 Mean Square Error

The Mean Squared Error (MSE) is another method classically used to measure a degree of likeliness between signals. It is defined as,

$$MSE = \frac{1}{N}(r(n) - x(n)^2)$$

3.3 Root Mean Square Error

$$RMSE = \sqrt{\frac{r(n) - x(n)2}{n}}$$

3.4 Normalized Root Mean Square Error (NRMSE)

$$NRMSE = \sqrt{\frac{[X(n)-r(n)]2}{[x(n)-\mu x(n)]2}}$$

estimate and the spectral subtraction as a filter, the filtered output is used not only for designing the filter but also as the input of the next iteration process.

3. Speech Objective quality measures

The objective comparison of three single channel speech enhancements is carried by evaluating performance of parameters such as, Mean Square Error (MSE), Normalized Mean Square Error (NRMSE), Signal to Noise Ratio (SNR), and Root Mean Square Error. It is based on mathematical comparison of the original and processed speech signal.

3.1 Signal to Noise Ratio (SNR)

It is most widely used and popular method to measure the quality of speech. It is ratio of signal to noise power in decibels.

$$SNR_{dB} = 10 \log_{10}((\sigma x)^2/(\sigma_d)^2)$$

Where $(\sigma_x)^2$ is the mean square of speech signal and $(\sigma_d)^2$ is the mean square difference between the original and

Where N is length of input speech signal, x(n) is input speech signal and r(n) is reconstructed speech signal.

4. Conclusion

In this paper, a comparison and simulation study of different forms of spectral subtractive-type algorithms for suppression of additive noise is presented. In particular, algorithms based on short-time Fourier transforms are examined and the limitations of spectral subtraction method are discussed briefly.

References

[1]Y. Ephraim, Statistical-Model-Based Speech Enhancement Systems, *The IEEE*, vol. 80, no. 10, pp. 1526–1555, October (1992).

[2] Sunil Devdas, "A multiband spectral subtraction method for speech enhancement" M.S.E.E, The University of Texas at Dallas, (2001).

[3] Dsphw – May 2008 lab-project 1: speech

Enhancement.
[4] Mohammad Reza Karami –Mollaei "Improvement on spectral subtraction method for speech enhancement", (IEEE Press, 2000).

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

[6] Y. Ephraim, H. L. Ari and W. Roberts, A Brief Survey of Speech Enhancement, The Electrical Engineering Handbook, 3rded. Boca Raton, FL: CRC, (2006).

[7] Y. Ephraim and I. Cohen, Recent Advancements in Speech Enhancement, The Electrical Engineering Handbook, CRC press, ch. 5, pp. 12–26, (2006).

[8]K.Yamashita,S.Ogata and T. Shimamura, Improved Spectral Subtraction Utilizing Iterative Processing, Electronics and Communications, Japan, Part 3, vol. 90, no. 4, pp. 39–51, (2007).

[10] Sheng Li, Jian-QiWang, Ming Niu, Xi-Jing Jing and Tian Liu, Iterative Spectral Subtraction Method for Millimeter-Wave Conducted Speech Enhancement, Journal of Biomedical Science and Engineering, vol. 3, no. 2, pp. 187–192, February (2010).

[11]Navneet Upadhyay and Abhijit Karmakar, Spectral Subtractive-Type Algorithms for Enhancement of Noisy Speech: An Integrative Review International Journal Image, Graphics and Signal Processing, vol. 5, no. 11, pp. 13–22, September (2013).