Jestr

Journal of Engineering Science and Technology Review 16 (4) (2023) 171 - 179

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

FPGA Implementation of Dynamic Quantile Tracking based Noise Estimation for Speech Enhancement

Bittu Kumar* and Ashwini Kumar Varma

Department of Electronics and communication Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad-500075, Telangana, India

Received 6 September 2022; Accepted 8 September 2023

Abstract

In this paper, the architecture of Dynamic Quantile Tracking (DQT) for estimating the noise signal has been proposed for Spartan-6 DSP Field Programmable Gate Array (FPGA) device. It is simulated on MATLAB/Simulink environment using Xilinx System Generator (XSG) blocks for speech enhancement application and implemented on Spartan-6 SC6SLX45 FPGA board in real-time mode. The hardware resources for the proposed architecture were obtained in terms of Flip Flops, Logic Slices, LUTs, Bonded IOBs, DSP48A1s and RAMB18E1s. The performance of the proposed architecture has been evaluated and analyzed with the variation of different speech-to-noise ratio levels and varying types of noises. Also, their results are compared with other methods like cascaded median and modified cascaded median methods. All results are evaluated and measured in teams of PESQ score with variations of SNR level and various noises and also presented in waveforms and spectrogram. In terms of PESQ score, it is observed that the enhanced speech for the proposed model is a good quality speech signal compared to other method's enhanced speech and noisy speech signals. In terms of hardware, it is observed that based on the utilization of resources, the proposed model acquires only 30% of the resources of the Spartan-6 SC6SLX45 FPGA board during real-time processing.

Keywords: Spectral Subtraction, Speech Enhancement, Dymanic Quantile Tracking based Noise Estimation, Simulink model

1. Introduction

Speech is the most practical and straightforward method for the human being to communicate their feelings and thoughts to others. Speech processing is required for human communication and in several applications such as mobile communication, hearing aids, multiparty teleconferencing and voice-controlled system. These systems perform well when there is no noise present in the environment, but their performance quickly degrades when there is noise in the working environment. Therefore, these systems' reliability needs to be improved. Speech enhancement technique gives a platform to reduce the noise level in speech signals. It is required for every application used for speech technologies, such as automatic speech recognition systems, speaker verification and speaker identification. To reduce the noises, speech enhancement algorithms are used in pre-processing stage of these technologies. Speech enhancement algorithms increase speech's perceived quality and intelligibility [1].

Due to available of microphones, for acquiring the speech data, it is grouped in different categories such as single-channel, dual-channel, or multi-channel speech enhancement techniques. In contrast to multiple-channel speech enhancement, which employs an array of numerous microphones, single-channel speech enhancement uses mainly one microphone. Dual-channel speech enhancement [2] uses mainly two microphones. Multi-channel speech enhancement techniques perform better than single-channel techniques or dual-channel techniques. Yet, single-channel techniques are more commonly used due to their simplicity and ease of hardware implementation. Also, single-channel speech enhancement techniques require less computation resources and memory consumption.

Based on the analysis, speech enhancement techniques can be implemented either in the time-domain or spectral domain. The spectral representation of the signal gives more information or features regarding the speech signal and noise. Therefore, spectral domain methods are more popular as compared to other domains. Several single-channel speech enhancement techniques [1-2] like as wiener filtering [3], MMSE STSA [4], spectral subtraction [5-7], signal subspace approach [8] and blind source separation [9] are available for obtaining the original speech signal from degraded speech signals. Among all, a few techniques-wiener filtering, MMSE and spectral subtraction methods are more popular.

The Wiener filtering [3] is a well-liked approach for improving speech quality that removes noise from speech signals. This filter minimizes the mean square error between the original and filtered signals. The power spectral density of the speech and noise components is determined by analyzing a noisy speech signal to use the Wiener filter for speech enhancement. Basically, the wiener filter requires a priori knowledge of the power spectra of the input speech signal, the noise, and the clean speech signal. Due to this, it cannot use if the speech signal and noise are non-stationary or non-Gaussian. The MMSE-based methods used the concept of minimum mean square error (MMSE) criterion [4]. It is based on minimum mean square error—short-time spectral amplitude (MMSE-STSA) i.e, *cost function* that minimizes the mean square error of the short-time amplitude spectrum. Through the literature survey, we investigated spectral subtraction methods have found lots of attention in the last few decades.

*E-mail address: kumarbittu135@gmail.com
ISSN: 1791-2377
doi:10.25103/jestr.164.21

In spectral subtraction, the noise (obtained by the noise estimation technique) [5] is subtracted from the degraded speech signal to produce the desired signal. Spectral subtraction is a popular technique for improving single-channel speech. It efficiently calculates the clean speech spectrum from the noisy speech spectrum by subtracting the approximate noise spectrum. The spectrum subtraction approach is carried out under two assumptions- 1) Speech and noise are assumed to be unrelated, 2) The phase of noisy speech signal is unaffected by noise. The spectral subtraction algorithm [6] has a trade-off between speech information and interference. It must be done to avoid speech distortion carefully. There is a possibility that we may lose some clean speech information while subtracting the noise from the noisy signal. If we subtract too much, some speech information may be lost; if too little is subtracted, much of the interfering noise (musical noise) may be present. Musical noise [7] is the noise with increasing variance that remains present in the estimated speech signal and may cause listening fatigue.

As we know that DSPs are specifically designed for the effective execution of typical DSP tasks, many engineers believe they are more energy-efficient than FPGAs. For instance, FPGAs can benefit from their highly parallel designs in some high-performance signal processing applications and provide substantially better throughput than DSPs. Therefore, despite DSP processors frequently having greater chip-level power consumption, the total energy consumption of FPGAs may be much lower than that of DSP processors. Hence, in this research work, we considered an FPGA-based approach for implementing the speech enhancement technique using Xilinx System Generator (XSG) blocks on Spartan- 6 DSP Field Programmable Gate Array (FPGA) device [16]. After implementation on Spartan-6, we computed the required hardware resources during the processing. In this technique, for estimating the noise spectrum, DQT is used, and then by the help of the spectral subtraction method, we obtained desired speech signal from the noisy speech signal. This paper also includes the simulation results in terms of PESQ with several SNR levels and noise types.

The rest of the paper is organized as follows. The speech enhancement technique has been defined in Section 1and the major section of this technique is DQT-based noise estimation, which is mentioned in Section 2. The Simulink model of complete speech enhancement technique, spectral subtraction method and DQT-based noise estimation are elaborated in section 3. Section 4 carried out detail about our experimental setup. The results in terms of PESQ score, hardware resources, waveforms, and spectrogram are discussed in section 5. Finally, section 6 concludes all the results of our paper.

2. Speech Enhancement Technique

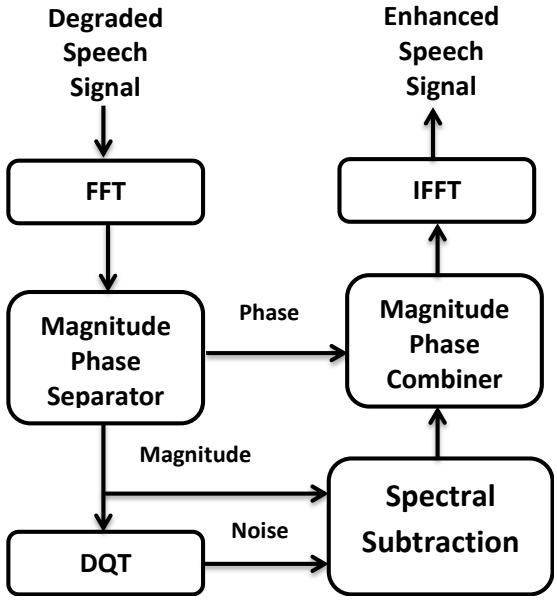

Fig. 1. Block diagram of speech enhancement technique using the spectral subtraction method

Figure 1 depicts the block diagram of the DQT noise estimation-based speech enhancement approach. This method uses Only noisy speech signals as input signals to find the original speech signals. The observed degraded speech signal, which includes both genuine speech data and background noise, is first transformed into a degraded speech spectrum $D(i,k)$ using the Fast Fourier Transform (FFT). The amplitude of the degraded speech spectrum is employed by the spectral subtraction approach to produce the clean speech spectrum $\widehat{C}(i,k)$. The generalized equation used for the spectrum subtraction approach is written as:

$$\left|\widehat{C}(i,k)\right| = \begin{cases} [|D(i,k)|^{\gamma} - a(N(i,k))^{\gamma}]^{1/\gamma}, & if\,|D(i,k)| > (a+b)^{\frac{1}{\gamma}}N(i,k) \\ b^{\frac{1}{\gamma}}N(i,k), & otherwise \end{cases} \quad (1)$$

where, $a = \text{over} - \text{subtraction factor}$,

$b = \text{floor factor}$ and

$\gamma = \text{exponent factor}$ (*Unity for magnitude spectral subtraction and two for power spectral subtraction method*)

Equation (1) calls for the noise spectrum $N(i,k)$, which is estimated using the noise estimation technique. The DQT-based noise estimation approach was selected since it requires less computing complexity and doesn't need memory. The true speech signal is obtained by removing the noise spectrum from the observed degraded speech spectrum, in accordance with the generalized equation of spectrum subtraction. A complex enhanced spectrum is created by combining the acquired spectrum (enhanced speech spectrum) and phase spectrum of the degraded speech signal. We used the Inverse Fast Fourier Transform (IFFT) to restore this spectrum into the time domain. In this approach, it was assumed that speech and noise were not adjusted and that noise did not affect the noisy speech spectrum's phase.

3. Literature Review

A crucial component of the speech enhancement method is noise estimation. The recovery of the excellent quality speech signal from the damaged signal is primarily a part of

the enhancement procedure. However, because the performance of the method typically depends on striking a balance between speech distortion and speech quality, it is challenging to eliminate noise without introducing distortion in the speech have been reported in the literature for finding the noise. The Voice Activity Detectors (VADs)-based technique [10] was extensively utilized to estimate noise. But, for low SNR-damaged speech signals or weak speech components, VAD does not track the absence of speech frames; instead, it identifies the silent structure in the degraded speech spectrum and periodically updates the noise. Another histogram-based method for estimating noise is described in [10-11]. It requires significant memory and fails when the SNR is low. Furthermore, since the signal segmentation takes place while building the histograms, which typically takes several hundred milliseconds, its noise estimation rate is acceptable. Useful noise estimation techniques, often called minimal tracking algorithms, have been published in recent publications. These methods assume that, even when speech signals are present, the power of the degraded speech spectrum frequently decays to the noise power. One may therefore measure the noise level by keeping an eye on the minimal power of the speech spectrum that has been degraded.

Several types of noise estimation methods have been availed in recently published papers by different authors. In most cases, researchers use statistical noise estimation methods [11]. Through the literature survey of noise estimation methods, we found that most researchers are more devoted to statistical noise estimation, the instant of VAD/histogram-based noise estimation, for computing noise in speech enhancement applications.

The noise estimation approach based on Minimum Statistics (MS) is one of the simplest tracking algorithms. In [12], the MS-based method tracks the minimum value of smoothed power of some previous noisy spectrum for estimating noise, followed by multiplying a factor that compensates for the bias. But, if the magnitude of the speech spectrum goes minima, then the minimum value of the speech spectrum is considered a noise which, however, results in loss of speech intelligibility, to some extent. Moreover, the MS-based method may sometimes eliminate low-energy phonemes, especially if the search window is too small. These limitations can be controlled, at the cost of significantly higher complexity, by including the smoothing parameter in time and frequency. This concept has been employed in Minima Controlled Recursive Averaging (MCRA) approach [14] for estimating the noise spectrum.

Authors [14] proposed the MCRA approach, in which averaging the previous power spectral values determines the noise spectrum. It also uses a smoothing parameter to adjust the sub-bands speech presence probability. Later, in [13], researchers proposed an Improved Minima Controlled Recursive Averaging (IMCRA) approach. The approach calculates the noise spectrum using a time-varying frequency-dependent smoothing parameter for balancing the speech presence probability controlled through minimum values of a smoothed periodogram. The drawback of minimal tracking algorithms is a very slow update when estimating the noise spectrum if there is a sudden change in the noise energy level.

In the quantile-based noise estimation method [15], Stahl et al observe that most of the frames (80-90%) of degraded speech carry a low energy level spectrum which is very near to the noise energy spectrum signal in the particular frequency bins and only a few numbers of frames (10-20%) of the signal contain high energy level spectrum corresponding to the speech signal. To this observation, noise samples are obtained as a few quantile values histogram of the observed degraded speech signals. More sorting operations are needed to compute the sample value at every quantile chosen in every frequency bin. In the cascaded median-based noise estimation method [16], Waddi et al reduce the sorting operation reduce in computational complexity and storage requirement. In [17-18], Kumar proposed a noise estimation, a modified version of cascaded median-based noise estimation, but it requires some storage requirements to improve speech quality. In [19-20], Dynamic Quantile Tracking (DQT) based noise estimation method does not require memory with less computational complexity.

4. DYnamic Quantile Tracking

With an increase or decrease in the estimated noise spectrum from the previous frame, we were able to dynamically calculate the quantile value for each individual frame in DQT-based noise estimation method. Changes between the estimated noise spectra of the current and previous frames may be discovered through the noise and degraded speech spectrum. Suppose, the magnitude of the degraded speech spectrum is larger than the previously estimated noise spectrum. In that case, the noise spectrum of the following frame will be increased by a tiny amount. In contrast, the noise spectrum of the next frame will be decreased if the size of the deteriorated speech spectrum is smaller than the previously calculated noise spectrum.

$$N(i,k) = \begin{cases} N(i-1,k) + \Delta_+(k) if |D(i,k)| \geq N(i-1,k) \\ N(i-1,k) + \Delta_-(k), \quad otherwise \end{cases} \tag{2}$$

Where,

$U(i,k)$ indicates noise spectrum,
$\Delta_+(k)$ denotes increment value of estimated noise spectrum
$\Delta_-(k)$ represents decrement value of the estimated noise spectrum respectively.

The condition for satisfying the ratio of increment and decrement of the estimated noise spectrum is followed as

$$\frac{\Delta_+(k)}{\Delta_-(k)} = \frac{q(k)}{1-q(k)} \tag{3}$$

Where,
$q(k)$is a quantile value in a particular frequency bin.

The following equations allow to choose the increment $\Delta_+(k)$ and decrement $\Delta_-(k)$.

$$\Delta_+(k) = \vartheta \times R \times q(k) \tag{4}$$

$$\Delta_-(k) = \vartheta \times R \times \big(1 - q(k)\big) \tag{5}$$

'R' indicates the range. It determines the difference between the highest and lowest spectral values. ϑ denotes a factor that is used to regulate the step size during tracking. If a quantile sample is over- or under-estimated as a result of an increase or decrease, the estimated ripple value may be computed as

$$\sigma = \Delta_+(k) + \Delta_+(k) = \vartheta \times R \quad (6)$$

Range is obtained by subtracting the peak $Pv(i,k)$ from valley$Vv(i,k)$. The peak and valley values are updated using the first order recursive equation, which is presented below equations:

$$Pv(i,k) = \begin{cases} \tau * Pv(i-1,k) + (1-\tau) * |N(i,k)|, & if|N(i,k)| \geq Pv(i-1,k) \\ \sigma * Pv(i-1,k) + (1-\sigma) * |Vv(i-1,k)|, & otherwise \end{cases} \quad (7)$$

$$Vv(i,k) = \begin{cases} \tau * Vv(i-1,k) + (1-\tau) * |N(i,k)|, & if|N(i,k)| \geq Vv(i-1,k) \\ \sigma * Vv(i-1,k) + (1-\sigma) * |Pv(i-1,k)|, & otherwise \end{cases} \quad (8)$$

$$R(i,k) = Pv(i,k) - Vv(i,k) \quad (9)$$

Where, σ and τ are constant factors used to control the fall and raise detection times, respectively. Figure 2 shows the flowchart of the DQT-based method. In this diagram, dynamic peak and valley detectors are worked for finding the peak and valley value according to equation 7 & 8, respectively. Range block calculates the range value using peak $Pv(i,k)$ & valley$Vv(i,k)$. By the help of R, we have the increment or decrement value of the estimated noise. Finally, from equation 2 we obtained the noise spectrum $N(i,k)$. The flowchart for the DQT-based technique is shown in Figure 2. In DQT, first, we take a degraded speech spectrum as input for calculating the peak and valley value of the degraded spectrum within the frame. Basically, for this task, dynamic peak detectors and dynamic valley detectors are used in this mythology to identify the peak and valley values in accordance with equations 7 and 8, respectively. After that, we find the range value by subtracting them as mentioned in equation (9).

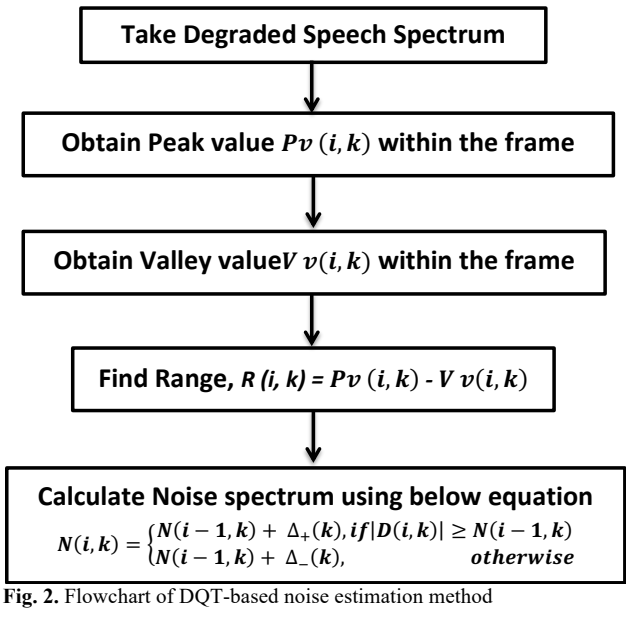

Fig. 2. Flowchart of DQT-based noise estimation method

5. Simulink Model for Speech Enhancement Techniquce

The DQT noise estimation-based speech enhancement technique using spectral subtraction is simulated using Xilinx System Generator (XSG) on MATLAB/Simulink platform and implemented on hardware Spartan- 6 DSP Field Programmable Gate Array (FPGA) device. First we have designed the architecture of a complete speech enhancement technique in MATLAB/ Simulink environment using the XSG blocks, which is shown in Figure 3. In this model, all blocks between *gateway In* block and *gateway Out* block, system generator and resource estimator are XSG blocks which came in MATLAB Library after the installation of ISE Xilinx software. Rests of the blocks are the internal blocks of Simulink Library. Some XSG blocks – FFT and CORDIC introduce delays in the simulation model, as shown in Table 1.

Table 1. Latency of Different XSG Blocks

Name of the XSG blocks	Latency (samples)
FFT 7.1.2	611
CORDIC 4.0.2	23
CORDIC 4.0.1	20
FFT 7.1.1 (IFFT)	611

In this model (*architecture*), major Xilinx System Generator (XSG) blocks are listed as follows -

- Fast Fourier Transform (FFT 7.1.2) – it use to transform the time domain noisy speech signal i.e. observed noisy speech signal (*signals come from Gateway In block*) into the frequency domain. The output of FFT i.e. noisy speech spectrum, comes in form of real value (*xk_re*) and imaginary value (*xk_im*)
- Magnitude and Phase (CORDIC 4.0.2) – it separates into magnitude part (*x_out*) and phase part (*phase_out*) from noisy speech spectrum (*it is in real and imaginary value*)
- DQT based noise estimation – it is used to find the noise spectrum with only requirement of single noisy speech spectrum. The separate architecture i.e. the figure 5 shows the simulation model of DQT based noise estimation method. According to the block diagram (*shown in figure 2*), the simulation model indicates every section such as range, peak detector, valley detector, and noise estimation. It gives noise spectrum only use of magnitude noisy spectrum (x_out).

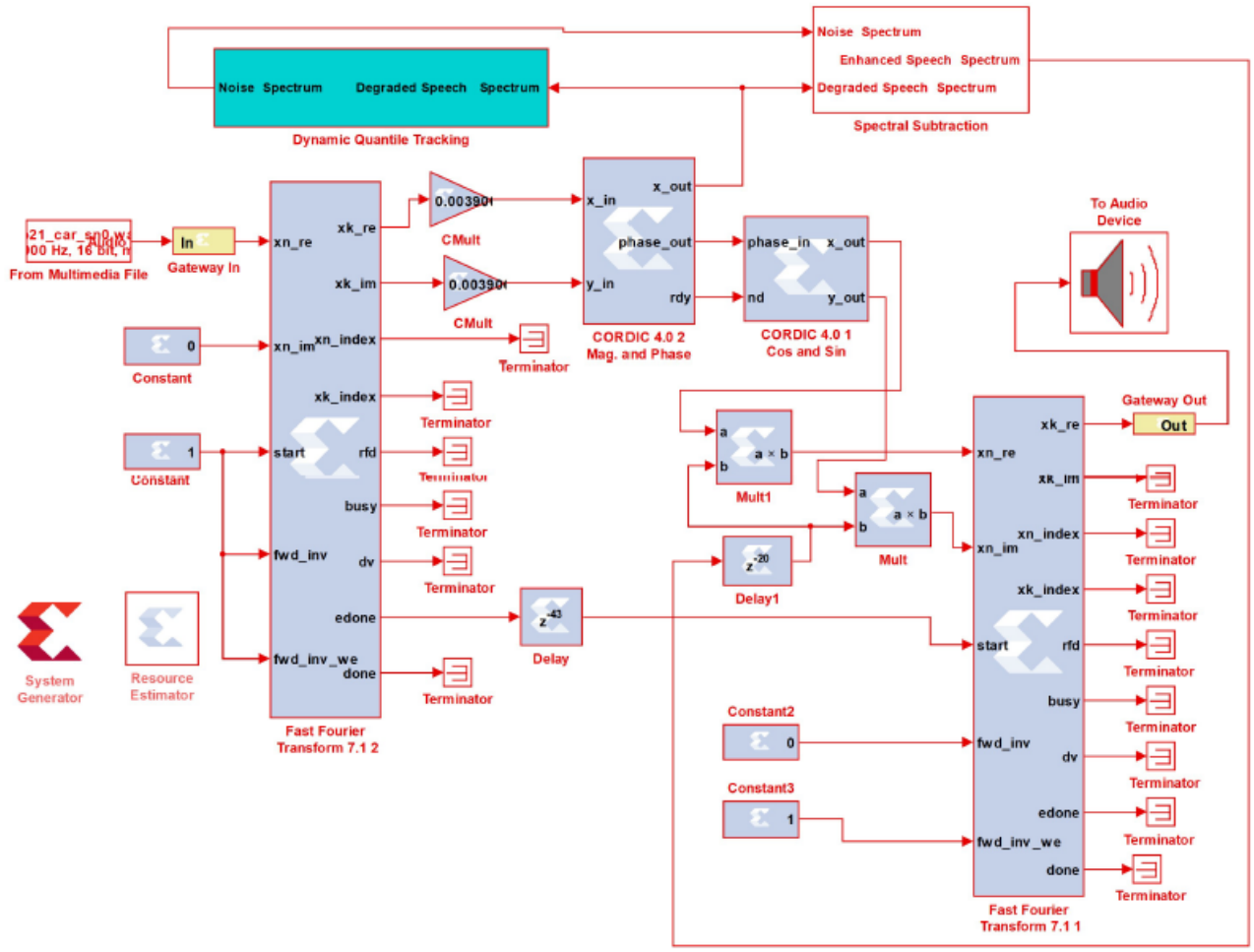

Fig. 3. Simulink model of complete speech enhancement using Xilinx System Generator (XSG) blocks

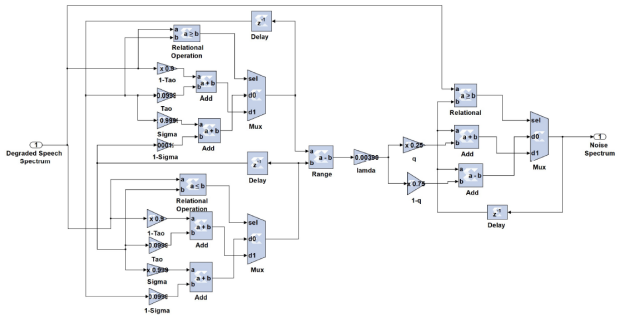

Fig. 4. Simulink model of Dynamic Quantile Tracking using Xilinx System Generator blocks

- Spectral Subtraction method – we have designed magnitude spectral subtraction by putting the unity value of the exponent factor in the generalized equation of spectral subtraction which is mentioned in equation 1. The architecture of spectral subtraction is shown in Figure 4. This architecture requires only noisy and noise spectrum for estimating enhanced speech spectrum, as the introduction shows.
- Real and Imaginary (CORDIC 4.0.1) – it is used to decompose the phase of the noisy speech spectrum (*phase_in*) into real value and imaginary value. These values – real (*x_out*) and imaginary (*y_out*) are multiplied separately with enhanced speech spectrum.
- Inverse Fast Fourier Transform (FFT 7.1.1) – it is applied to transform into time domain speech signal (*xk_re*) from enhanced speech spectrum, which is in real (*xn_re*) and imaginary (*xn_im*).

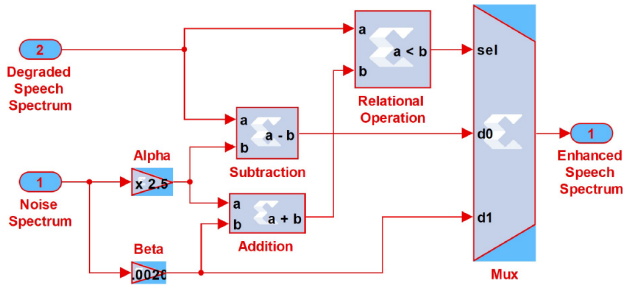

Fig. 5. Simulink model of Spectral Subtraction using Xilinx System Generator (XSG) blocks

6. Experimental Setup

In this section, we organized an experimental setup to test the performance of the proposed architecture for DQT-based noise estimation in real-time mode. The image of our setup is shown in Figure 6. In this setup, we used Spartan-6 SC6SLX45 FPGA board [21-22] and Laptops with installed Xilinx ISE Design Suite 14.7 software. First, we have created a simulation model through the Simulink model (which was made using Xilinx System Generator blocks

only) and obtained simulation results discussed in the results section. We obtain a hardware Co-simulation block using the system generator token in the MATLAB platform [23]. This hardware co-simulation block is replaced by all XSG blocks which are placed in between Gateway In block and Gateway Out block. Here, it can say that the hardware co-simulation block represents whole blocks which are present in between Gateway In and Gateway Out block. We connect the FPGA device with laptops and load the bit-stream file (it was generated together with the Co-simulation block). After running the simulation model, we got the enhanced speech signal.

Fig. 6. Experimental Setup (PC with FPGA device)

7. Results and Discussion

7.1. Parameter Setting

The architecture of DQT-based noise estimation has been simulated in MATLAB(R) R2011b environment. For testing of the proposed model, degraded speech signals are selected from SpEAR (Speech Enhancement Assessment Resource) database, these speech file are recorded separately in both male and female speakers with different SNR levels as well as with different noises. In our experimentation, speech qualities of enhanced speech signals have evaluated in terms of PESQ scores and waveforms. Also, DQT based simulation model for speech enhancement has been implemented on the Spartan-6 SC6SLX45 FPGA board. Hardware resources of this model have been estimated. Before the simulation, the values of parameters of the spectral subtraction method and DQT-based noise estimation method are applied as given in Table 2.

Table 2. Parameter Setting

Name of the Parameters	Parameter's value
Over-subtraction factor (a)	2.5
Spectral floor factor(b)	0.002
Exponent factor(γ)	1
Factor 'τ'	0.1
Factor 'σ'	$(0.9)^{1/1024}$
Convergence factor (λ)	1/256
Quantile value (p) (initial value)	0.25

In among the value of all parameters, the quantile value (p) is an essential role play in noise estimation. As in [15], the same value of p were taken in our experimentation. We have organized an experiment to justify the reason behind selecting the same value of$p$. In this experiment, three degraded noisy speech files were taken, and it passed through our proposed model with different quantile values (p). The obtained enhanced speech signals find their PESQ score using clean speech signal, which has also taken the SpEAR database. Figure 7 shows the PESQ score of enhanced speech signal for different quantile values. From this figure, it is cleanly seen that the PESQ score increases with a quantile value up to 0.3. From our observations, in the range$p = 0.2 - 0.3$, the proposed model gives the highest PESQ score or better speech signal. We select the middle value i.e. 0.25, for the quantile value.

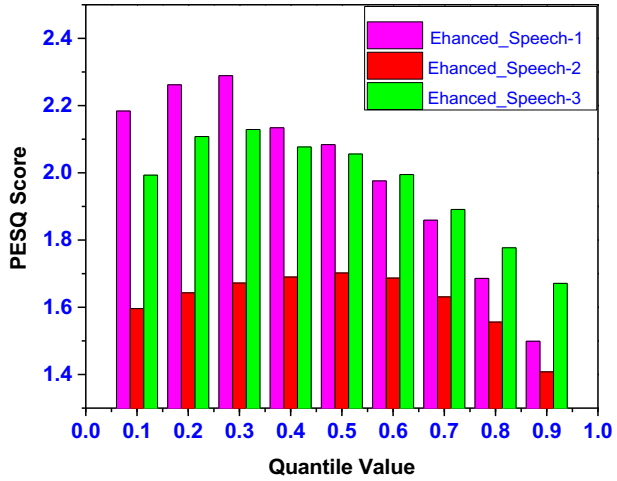

Fig. 7. Quantile Value vs PESQ score

7.2. Resource Utilization

Table 3 shows the resource utilization of the complete speech enhancement technique's Real-time Simulink model (*shown in Figure 3*). The values of resources such as flip flops, LUTs (*Look Up Tables*), logic slices (*each slices carries eight flip flops and four LUTs*), Bonded IOBs, DSP48A1s (*it contains adder, 18×18 multiplier and accumulator*) and RAMB18E1s (*18Kb Block RAM*) are seen in the obtained report that is generated through '*resource estimator*' block in Simulink model by Xilinx ISE Design Suite. Table-III shows that the percentage of utilization of resources (or used resources) is less than 30%, except for DSP48A1s.

Table 3. Report on Resources Utilization

SL No.	Name of Resources	Used	Available	Utilization (in %)
1.	Flip Flops	7,492	54,576	13.73
2.	Logic Slices	2,042	6,822	29.93
3.	LUTs	6,774	27,288	24.82
4.	Bonded IOBs	34	228	14.91
5.	DSP48A1s	28	58	48.28
6.	RAMB18E1s	10	116	8.62

7.3. PESQ Results

The proposed DQT-based noise estimation method model has been tested using objective measures i.e. PESQ score. From the SpEAR database [24], we have chosen 28 different degraded speech signals corrupted by different noises (*burst, f16, factor, pink, Volvo and white*) with different SNR levels such as 1, 3, 5, 7, 9, 11, 13 and 15dB. These speech files (*spoken by both males and females*) pass through the proposed speech enhancement model and store the enhanced speech signals. Then, we calculate the PESQ score of enhanced speech signals as well as degraded speech signals. Figure 8a shows the simulation result (*plot between PESQ*

score vs SNR)for two different speech files – "noisy speech file 1*"*(*spoken by male*) and "noisy speech file 2*"*(*spoken by female*)which are corrupted with f16 cockpit noise and pink noise respectively. In this figure, the PESQ score [25] of recovered speech signals (*enhanced speech signals*) is higher as compared to observed speech signals (*noisy speech signals*) at each SNR level - 1, 3, 5, 7, 9, 11, 13 and 15dB. The plot of different Noises vs PESQ scores is shown in Figure 8b. This plot contains the PESQ score for the same noisy speech files (*noisy speech file 1 and noisy speech file 2*), which was taken in Figure 8a but the change in SNR level and types of noises. In this case, we have taken different noisy speech files corrupted with different noises (*burst, f16, factor, pink, Volvo and white*) at 1 dB SNR level. From the figure, it observed that the PESQ value of enhanced speech is more significant than noisy speech for each noise.

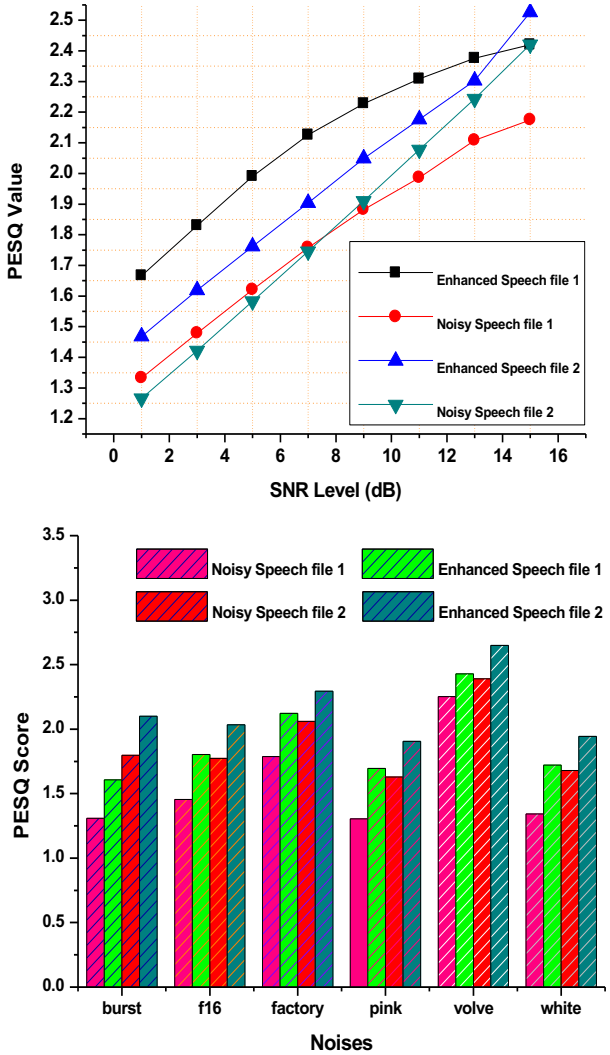

Fig. 8. (a) SNR vs PESQ score (b) Noises vs PESQ score for two different noisy speech signals and enhanced speech signals

In the above experiment (*results shown in Figure 8*), we observed that enhanced speech by the proposed model has good quality compared to the noisy speech signal under different conditions like SNR variation and the case of varying noise types. To find a more conclusive statement about the performance of the proposed model with the comparison of different noise methodologies (Cascaded median and modified cascaded median-based noise estimation methods), we need to perform more experiments. For this, six noisy speech files select from the same database i.e. SpEAR database. These speech files individually are corrupted by six different noises – burst, f16, factory, pink, volvo and white noise; every speech file carries the same sentence "*good service should be rewarded by big tips*". Figure 9 shows the comparison results of enhanced speech signals in terms of PESQ score for different noise estimation methods. As seen from Figure 9, out of six enhanced speech signals, the PESQ score of three enhanced speech signals (in F16, factory and volvo noise cases) for the proposed architecture is superior as compared to the other two methods i.e Cascaded median and modified cascaded median based noise estimation methods.

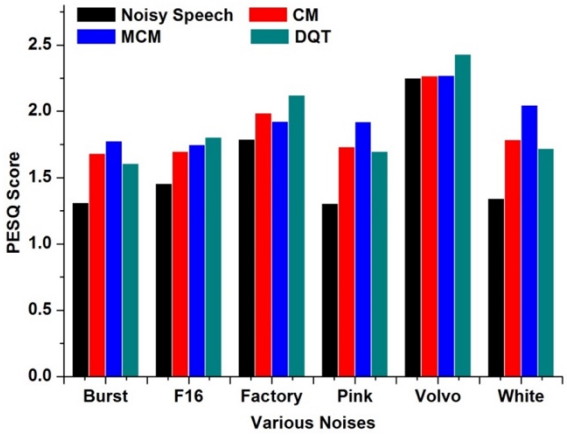

Fig. 9. Noises vs PESQ score for two different noisy speech signals and enhanced speech signals.

7.4. Waveforms and Spectrogram

This section presents the visualization of clean/original speech, degraded speech and enhanced speech signal in waveform and spectrogram. The degraded speech and its original speech signal are chosen from same database i.e, the SpEAR database. The selected degraded speech signal contains the sentence *"butterscotch fudge goes well between your ice-cream"*. It is store in wave format with 16 KHz sampling frequency. It is recorded by a male voice under the pink noise environment. Figures 10, 11 and 12 show the waveform and spectrogram of clean, noisy, and enhanced speech signals, respectively. From these figures, it can be observed that Figure 11 and 12 are very close to each other; i.e., enhanced speech signal is perceptually good in hearing by normal listeners.

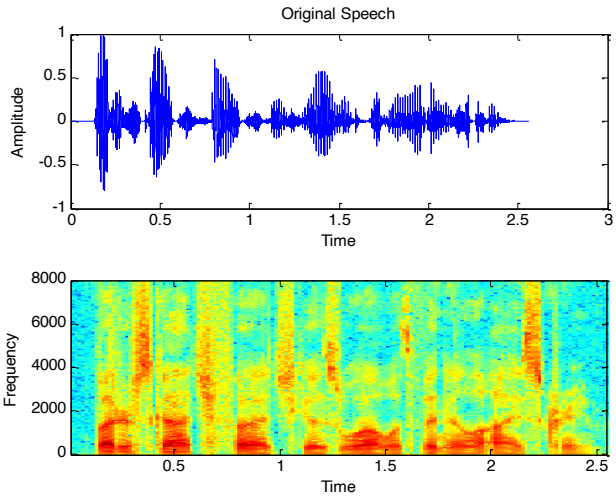

Fig. 10. Waveform and spectrogram for original speech Signal

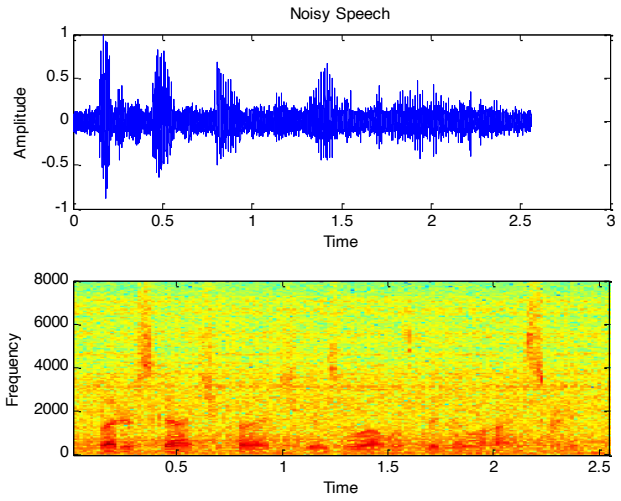

Fig. 11. Waveform and spectrogram for degraded speech (pink noise at 10 dB)

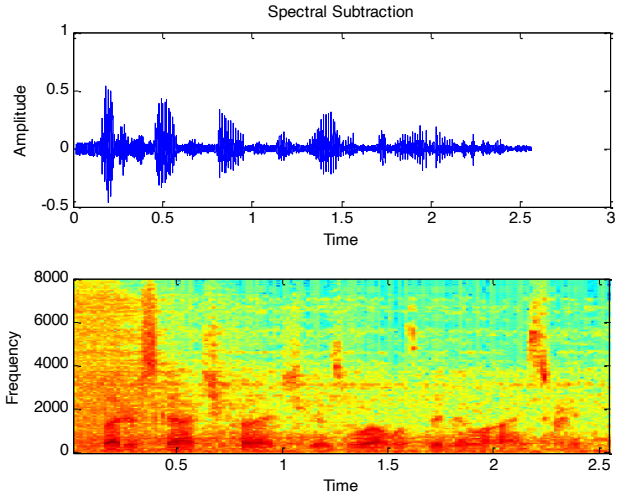

Fig. 12. Waveform and spectrogram of enhanced speech by DQT

8. Conclusion

This paper has proposed the architecture of DQT- based noise estimation method. The proposed architecture has been tested for speech enhancement using spectral subtraction and successfully implemented on the Spartan-6 SC6SLX45 FPGA board. Also, its performance has been tested with different degraded speech signals in real-time process. Results in terms of PESQ score and waveforms have been reported in this paper. Further, the proposed architecture results have been compared with Cascaded median and modified cascaded median-based noise estimation methods. Among all analyses of simulation results, especially in terms of PESQ score, it is observed that the enhanced speech for the proposed model is a good quality speech signal compared to other method's enhanced speech and noisy speech signals. In terms of hardware, the Essential resources of the proposed model are estimated for implementation in real-time processing on FPGA devices. It is observed that based on the utilization of hardware resources, the proposed model acquires only 30% of the resources of the Spartan-6 SC6SLX45 FPGA board during real-time processing. In the future, it may also investigate the implementation of the proposed model using other FPGA devices and evaluate its performance with different speech corpus.

References

[1] Chaudhari and S. B. Dhonde, "A review on speech enhancement techniques," in *2015 International Conference on Pervasive Computing (ICPC)*, Pune, India: IEEE, Jan. 2015, pp. 1–3. doi: 10.1109/PERVASIVE.2015.7087096.

[2] I. Kaur, V. K. Nassa, T. Kavitha, P. Mohan, and S. Velmurugan, "Maximum likelihood based estimation with quasi oppositional chemical reaction optimization algorithm for speech signal enhancement," *Int. j. inf. tecnol.*, vol. 14, no. 6, pp. 3265–3275, Oct. 2022, doi: 10.1007/s41870-022-01032-6.

[3] Jaiswal, R.K., Yeduri, S.R. and Cenkeramaddi, L.R, "Single-channel speech enhancement using implicit Wiener filter for high-quality speech communication," *Int J Speech Technol,* vol.25, no.3, pp.745–758, Sept.2022, doi:10.1007/s10772-022-09987-4.

[4] Kumar, B. "Comparative performance evaluation of MMSE-based speech enhancement techniques through simulation and real-time implementation," *Int J Speech Technol,* vol. 21, no.04, pp.1033–1044, Dec. 2018, doi: 10.1007/s10772-018-09567-5.

[5] Abajaddi, N., Mounir, B., Elmaazouzi, L., Mounir, I., Farchi, A, "Speech Spectral Subtraction in Modulation Domain," *Lecture Notes in Networks and Systems*, vol.357, pp. 239-246, Jan. 2022, doi: 10.1007/978-3-030-91738-8_23.

[6] Kumar B, "Mean-Median based Noise Estimation Method using Spectral Subtraction for Speech Enhancement Technique," *Indian J. Sci. Technol*, Vol. 9, no. 35, pp:1-6, Feb. 2016, doi: 10.17485/ijst/2016/v9i35/100366.

[7] M. Bahoura and H. Ezzaidi, "Implementation of spectral subtraction method on FPGA using high-level programming tool,"*24th International Conference on Microelectronics (ICM)*, Algiers, Algeria, IEEE, Dec. 2012, pp. 1-4, doi: 10.1109/ICM.2012.6471434.

[8] Sun, Chengli, JianxiaoXie, and Yan Leng. "A Signal Subspace Speech Enhancement Approach Based on Joint Low-Rank and Sparse Matrix Decomposition," *Archives of Acoustics,* vol.41, no. 2, pp: 245-254. Jan. 2016, doi: 10.1515/aoa-2016-0024.

[9] Kumar, B "Comparative performance evaluation of greedy algorithms for speech enhancement system," *Fluctuation and Noise Letters,* vol. 20, no. 02, Oct. 2021, doi: 10.1142/s0219477521500176.

[10] M. Oukherfellah and M. Bahoura, "FPGA implementation of voice activity detector for efficient speech enhancement," *IEEE 12th International New Circuits and Systems Conference (NEWCAS)*, Trois-Rivieres, QC, Canada, Oct 2014, pp. 301-304, doi: 10.1109/NEWCAS.2014.6934042.

[11] H. M. Goodarzi and S. Seyedtabaii, "Speech Enhancement Using Spectral Subtraction Based on a Modified Noise Minimum Statistics Estimation," *2009 Fifth International Joint Conference on INC, IMS and IDC*, Seoul, Korea (South), Nov. 2009, pp. 1339-1343, doi: 10.1109/NCM.2009.272.

[12] Kum, Jong-Mo, Yun-Sik Park, and Joon-Hyuk Chang "Improved minima controlled recursive averaging technique using conditional maximum a posteriori criterion for speech enhancement", *Digital Signal Processing*, vol.20, no. 6, pp: 1572-1578, Dec. 2010, doi: 10.1016/j.dsp.2010.01.011.

[13] Seyedtabaee, S., Moazami Goodarzi, H. Improved Noise Minimum Statistics Estimation Algorithm for Using in a Speech-Passing Noise-Rejecting Headset. *EURASIP J. Adv. Signal Process.* , vol. 395048, June 2010, doi: 10.1155/2010/395048

[14] N. Fan, J. Rosca and R. Balan, "Speech Noise Estimation using Enhanced Minima Controlled Recursive Averaging," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Honolulu, HI, USA, June 2007, pp. IV-581-IV-584, doi: 10.1109/ICASSP.2007.366979.

[15] V. Stahl, A. Fischer and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, Istanbul, Turkey, Aug. 2000, pp. 1875-1878, doi: 10.1109/ICASSP.2000.862122.

[16] S. K. Waddi, P. C. Pandey and N. Tiwari, "Speech enhancement using spectral subtraction and cascaded-median based noise estimation for hearing impaired listeners," *2013 National Conference on Communications (NCC)*, New Delhi, India, Feb. 2013, pp. 1-5, doi: 10.1109/NCC.2013.6487989.

[17] Kumar, B, "Real-time performance evaluation of modified cascaded median-based noise estimation for speech enhancement system," *Fluctuation and Noise Letters,* vol. 18, no. 04, April 2019, doi: 10.1142/S0219477519500202.

[18] Kumar, B.“ Spectral Subtraction using Modified Cascaded Median based Noise Estimation for Speech Enhancement,” *Sixth International Conference on Computer and Communication Technology (ICCCT 2015)*, MNNIT, Allahabad, India, Sept. 2015. pp. 214-218, doi: 10.1145/2818567.2818608.

[19] Tiwari, N., Pandey, P.C. A “Technique with Low Memory and Computational Requirements for Dynamic Tracking of Quantiles,” *J Sign Process Syst, vol.* 91, no. 05, pp. 411–422, May 2019, doi: 10.1007/s11265-017-1327-6.

[20] N. Tiwari and P. C. Pandey, "Speech Enhancement Using Noise Estimation With Dynamic Quantile Tracking," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2301-2312, Dec. 2019, doi: 10.1109/TASLP.2019.2945485.

[21] Ian Kuon, Russell Tessier and Jonathan Rose, "FPGA Architecture: Survey and Challenges," *Foundations and Trends® in Electronic Design Automation*, Vol.02: No.02, April 2008, pp 135-253, doi: 10.1561/1000000005

[22] Vivekanand and B. Kumar, "Melody extraction in noisy Musicof different genre," *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)*, Bhopal, India, Oct. 2017, pp. 629-632, doi: 10.1109/RISE.2017.8378229.

[23] Spartan-6 FPGA Configuration User Guide (UG380), www.xilinx.com/support/ *documentation/user guides/ug380.pdf*

[24] Speech Enhancement Assessment Resource (SpEAR) Database. http://ee.ogi.edu/NSEL/. Beta Release v1.0. CSLU, Oregon Graduate Institute of Science and Technology. E. Wan, A. Nelson, and Rick Peterson..

[25] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Salt Lake City, UT, USA, May 2001, pp. 749-752 vol.2, doi: 10.1109/ICASSP.2001.941023.