Design of FPGA-Based Rapid Prototype Spectral Subtraction for Hands-free Speech Applications

Sarayut Amornwongpeeti*, Nobutaka Ono*, and Mongkol Ekpanyapong†
*Principles of Informatics Research Division, National Institute of Informatics, Tokyo, Japan
E-mail: sarayut.amornwongpeeti@ait.ac.th, onono@nii.ac.jp Tel: +81-3-4212-2827
†Department of Microelectronics and Embedded Systems, Asian Institute of Technology, Pathum Thani, Thailand
E-mail: mongkol@ait.ac.th Tel: +662-524-5689

Abstract— **In this paper, the design of a FPGA-based rapid prototype for Short-Time Fourier Transform (STFT) based spectral subtraction for hands-free speech applications using Xilinx System Generator (XSG) tools without traditional HDL hand coding is presented. Initially, the concept of a dual-channel short-time spectral subtraction algorithm for removing the wideband background noise in a speech signal is introduced. The studied algorithm is developed in the system-level modeling simulator using MATLAB Simulink environment. For the digital hardware design, simple hardware architectures for data framing and overlapping algorithms are proposed by utilizing the basic DSP blocksets of the XSG library, and replaced in the simulation model. Finally, a complete simulation model of the FPGA-based short-time spectral subtraction algorithm using XSG is presented. The comparative performance evaluation based on simulation results and the summary of resource utilization are confirmed the implementation feasibility of the real-time FPGA-based short-time spectral subtraction algorithm for hands-free speech applications.**

I. INTRODUCTION

For the application domain in the area of audio and acoustic signal processing, it is inevitable that the input speech signal captured by the microphone sensor is always mixed with the background noise in real-world environments. Depending on the level of the ambient noise, it can significantly degrade the desired speech signal in term of the speech quality and intelligibility. However, the presence of additive acoustic noise in the interested speech can be reduced and alleviated by using a broad range of many existing speech enhancement techniques, such as spectral subtraction [1,4], signal-subspace embedding [2], time-domain iterative approaches [3]. The main objective of these speech enhancements is to remove the background noise while maintain the perceptual aspects near the level of the noise-free input speech.

Spectral subtraction is a simple and efficient single-channel speech enhancement technique, which was originally proposed in the STFT domain by [1]. Later, the spectral subtraction approach with a broad class of estimators including the magnitude and the power subtraction has been studied in [4]. The concept of this method is to suppress the additive acoustic noise from the noisy speech by subtracting either the magnitude or the power spectrum between the observed signal and the estimated noise spectrum. For the hands-free speech applications, multiple input microphones of the multi-channel speech system can be placed far away from the speaker of interest [5]. In this paper, the system is considered as a dual-channel speech enhancement, where two signal inputs are available making the system possible to use the auxiliary channel as the noise reference as shown in Fig. 1. In addition, it is assumed that the background noise source is located far from these two microphone sensors so that the non-stationary noise variance of two channels can be considered in equal except the noise waveform in time domain. By this way, adaptive noise cancellation can be easily achieved only by placing the reference microphone far away from the speaker of interest.

Nowadays, the hardware-based FPGA platform has been recognized as a promising technology for the high performance real-time digital signal processing system. FPGAs offer promising features over than the traditional DSPs including, fast computation time due to the highly internal parallel processing, customizable platform with design flexibility, and high reliability. By using hardware description languages, such as VHDL and Verilog HDL, complete digital logic systems can be simply designed in hierarchical modules and synthesized into the FPGA. Xilinx System Generator (XSG) is a high-level graphical programming tool used for designing a high performance DSP system with MATLAB Simulink targeting for the FPGA. XSG tool allows the hardware designer to build a system-level model in Simulink for algorithm verification, such as adaptive noise canceling [6,7], STFT based pattern recognition [8], spectral subtraction [9-12]. In addition, the XSG code generator also enables the user to automatically generate the synthesizable HDL code mapping into the highly pre-optimized IP cores, which dramatically reduced the design time and quickly evaluate new algorithms in hardware comparing with traditional HDL hand coding.

According to the literature, although the power spectral subtraction architecture using high-level XSG tool has been presented in [9], the absence of hardware design for the DFT framing-windowing algorithm as well as the inverse DFT overlap-add algorithm leads to the algorithm limitation for the real FPGA hardware implementation. In [10,11], a complete

978-616-361-823-8

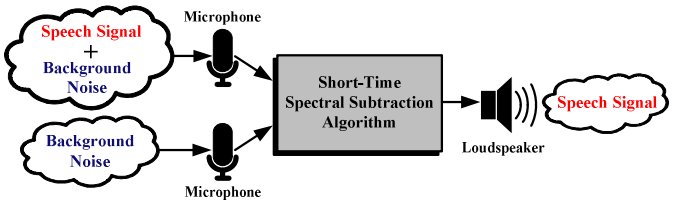

Fig. 1 Overall system of dual-channel short-time spectral subtraction for hands-free speech applications.

FPGA-based system of the magnitude spectral subtraction with the XSG design approach has been successfully implemented. However, the use of only initial data frames for estimating the entire noise magnitude spectrum make the system less attractive for hands-free speech applications, which often involves the non-stationary noise in the real environment. The initial frame must be contained only the noise component in order to estimate the average constant noise floor. Although the data framing and the overlap-add operations were implemented with the basic Xilinx addressable shift register block, this design is limited with the maximum memory depth of 1024 resulting in the limitation for the large data buffer size. A FPGA-based magnitude spectral subtraction using the traditional HDL coder-based approach has been developed in [12]. However, this approach is not suitable for the rapid prototype development due to large time consuming during the design process.

Thus, the aim of this paper is to present the design of a FPGA-based rapid prototype short-time spectral subtraction for hands-free speech applications using XSG tools and also propose simple hardware architectures for FPGA implementation of the data framing and overlapping algorithms, which can be expandable for the large data buffer size. The outline of this paper is organized as follows. The introduction is briefly discussed in Section 1. In Section 2, the concept of short-time spectral subtraction for reducing the ambient acoustic noise is described. In Section 3, the system-level simulation model and the XSG-based architecture model of the studied algorithm is developed and presented. In Section 4, simple hardware architectures utilizing the basic XSG blocksets for data framing and overlapping algorithms is proposed. Finally, simulation results and conclusions are covered in Section 5 and Section 6, respectively.

II. The Concept of Short-Time Spectral Subtraction

The principles of short-time spectral subtraction is considered based on the transform domain approach, which consists of the short-time analysis-modify-synthesis steps in the frame-based processing. Based on the assumption that the noise is an additive zero-mean and uncorrelated with the clean speech in the wide-sense stationary random process. Thus, the observed noisy signal $x(t)$ is a sum of the clean speech signal $s(t)$ and the additive white noise $d(t)$, which commonly given as:

$$x(t) = s(t) + d(t) \tag{1}$$

A. Short-time Fourier Transform

The short-time Fourier transform (STFT) is two-dimensional representation in the meaningful time and frequency domains. The STFT of a signal $x(t)$, evaluated at time index m and frequency index n is defined as [13]:

$$X(m,n) = \sum_{t=mR}^{mR+N-1} w(t-mR)x(t)e^{-j2\pi(t-mR)n/N} \tag{2}$$

whereas the inverse STFT of the processed signal $Y(m,n)$ is:

$$y(t) = \sum_{m=-\infty}^{+\infty} w(t-mR)\left(\frac{1}{N}\sum_{n=0}^{N-1} Y(m,n)e^{j2\pi(t-mR)n/N}\right) \tag{3}$$

where N is the size of frame length, R is the frame shift factor, $w(t)$ is the windowing function. For perfect reconstruction,

$$\sum_{m=-\infty}^{+\infty} w(t-mR)^2 = 1 \tag{4}$$

should be satisfied.

B. Short-time Power Spectral Subtraction

By considering the stereo-channel speech enhancement system, let us denote $X_1(m,n)$ as the STFT representation of the first channel containing the target speech and the background noise, while $X_2(m,n)$ is the STFT representation of channel two containing only the background noise. By using the STFT analysis into (1), the mathematical expression of mixed model in STFT domain can be written as:

$$X_1(m,n) = S(m,n) + D(m,n) \tag{5}$$

Thus, the STFT magnitude squared for power spectral subtraction method can be expressed as [4]:

$$\begin{aligned}|X_1(m,n)|^2 = {} & |S(m,n)|^2 + |D(m,n)|^2 + S(m,n)D^*(m,n) \\ & + S^*(m,n)D(m,n)\end{aligned} \tag{6}$$

Due to the uncorrelated noise with the clean speech signal, the cross term between these two signals can be statistically set to zero, which can be simplified and rearranged as:

$$|\hat{S}(m,n)|^2 = |X_1(m,n)|^2 - |D(m,n)|^2 \tag{7}$$

where $|\hat{S}(m,n)|^2$ and $|D(m,n)|^2$ are the estimated speech and noise power spectrum, respectively. In [13], the real positive gain $G(m,n)$ is defined as:

$$G(m,n) \simeq \frac{|\hat{S}(m,n)|}{X_1(m,n)} \tag{8}$$

By substituting in (7), the defined gain ($0 < G(m,n) < 1$) can be rewritten as:

$$\begin{aligned}G(m,n) &\simeq \frac{\sqrt{|X_1(m,n)|^2 - |D(m,n)|^2}}{X_1(m,n)} \\ &\simeq \sqrt{1 - \frac{1}{Q(m,n)}}\end{aligned} \tag{9}$$

where the relative signal level $Q(m,n)$ can be defined as:

$$Q(m,n)=\frac{|X_1(m,n)|^2}{|D(m,n)|^2} \tag{10}$$

For avoiding the spectral subtraction estimates negative spectral magnitudes, the value of $Q(m,n)$ is replaced by one if it is less than one. Finally, the processed signal $Y(m,n)$ can be expressed as a function of the real positive gain $G(m,n)$ and obtained as:

$$Y(m,n)=G(m,n)\cdot X_1(m,n) \tag{11}$$

C. Noise Estimation

In general, the noise estimation can be determined by either the assumed known properties of the background noise or the actual measurement during the intervals of the input speech absence [4]. In this paper, for the simplicity it is assumed that the acoustic ambient noise is accessible on a separate microphone channel such as in hands-free speech applications. Thus, the effect of additive background noise on the signal spectrum can be easily estimated by taking the data from the accessible auxiliary microphone. For estimating noise power spectrum $|D(m,n)|^2$, we can use the other observation channel. Let $X_2(m,n)$ be the STFT representation of the other channel. If we can assume that this channel captures only noise, we can estimate as:

$$|D(m,n)|^2=\beta|X_2(m,n)|^2 \tag{12}$$

where β is a control parameter to compensate the sensitivity mismatch between two channels. By this way, the actual noise variance with different variations can be evaluated accurately and adaptively suppressed in the real-time processing. Fig. 2 shows the basic structure of dual-channel short-time spectral subtraction for hands-free speech applications.

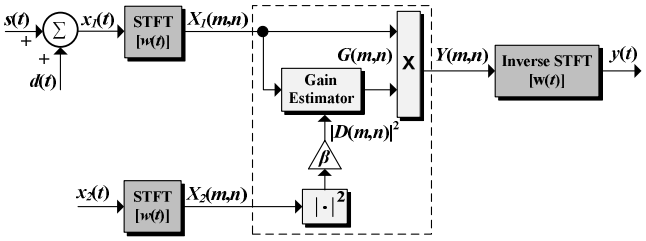

Fig. 2 Basic structure of dual-channel short-time spectral subtraction.

III. Simulation Model of Short-time Spectral Subtraction

A. System-level MATLAB-based Model

Initially, for algorithm verification the short-time spectral subtraction algorithm is design and simulated at the system-level model with the floating-point MATLAB implementation in Simulink environment. The MATLAB-based model mainly consists of block diagrams of the STFT, the gain estimator, and the inverse STFT as shown in Fig. 3, which developed based on the mathematical equations from (1) to (12).

B. System-level XSG-based Model

As mentioned in the previous section, the XSG toolbox does not only offer a high-level graphical programming tool that can be simulated and then configured to the FPGA target but also an efficient solution for FPGA-based rapid prototype development. In contrast to HDL coder approach, the system verification can be easily made and explored in several algorithm variations with XSG model-based approach. However, the one-to-one conversion between the Simulink functional blocks and the XSG blocksets has to be made manually. In this paper, the basic XSG blocksets for DSP are

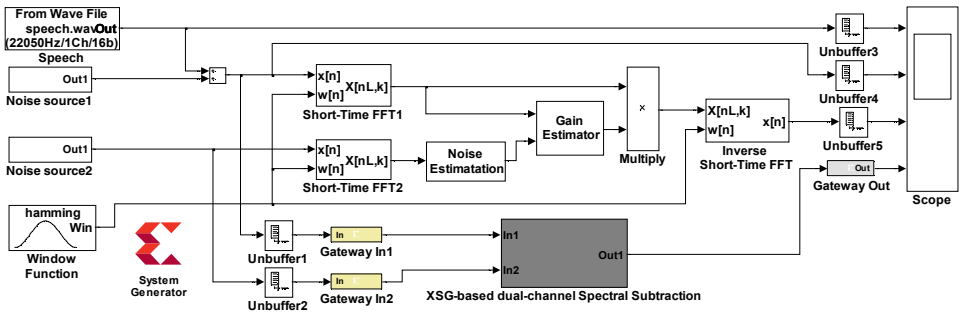

Fig. 3 Top-level simulation diagram showing a MATLAB-based model and a XSG-based model of short-time spectral subtraction.

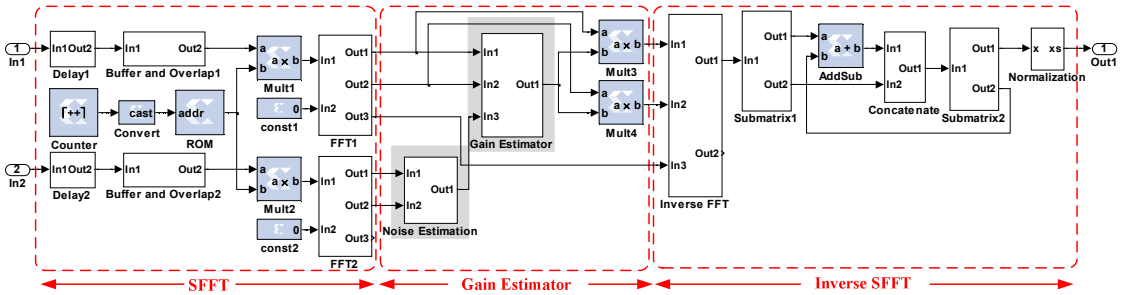

Fig. 4 XSG-based hardware architecture of dual-channel short-time spectral subtraction.

used for the hardware design and replaced in the MATLAB-based simulation model. Several XSG library blocksets, including a wide range of primitive basic blocksets, Xilinx FFT block, CORDIC divider, and CORDIC SQRT, are utilized for the XSG-based fixed-point implementation model for a high performance DSP system of the short-time spectral subtraction algorithm as shown in Fig. 4.

IV. XSG Design and FPGA Implementation

A. Short-time Fourier Transform Module

The STFT module consists of Hamming window, buffer with overlap, and FFT subsystem. The Hamming window is implemented with the Xilinx single port ROM with 512 depths and an address count-limited counter. Since the provided XSG library is still limited to some basic functional blocksets, which cannot be achieved for the sophisticated signal processing functions and complex operations, such as framing and overlap-add functions. In this paper, the data framing is implemented with the proposed architecture including a selected multiplexer, 8 identical delay lines (a cascaded multiplexer with unit delays), and a control logic circuit as shown in Fig. 5. The buffer size and overlap are set to 512 and 256, respectively. The Xilinx FFT block is used to perform the FFT function with 512 transform length, which provides two output signals of the real and imaginary parts. For continuously computation of STFT, the FFT pipelined streaming I/O option is selected for hardware implementation.

B. Gain Estimator Module

In the short-time Fourier transform domain, the gain is determined individually at each frequency region and apply to each frequency bin of the input noisy speech. In this paper, the gain estimator module consists of the power spectral calculation and the amplitude limiter, which can be easily implemented with the pre-optimized IP cores. Fig. 6 shows the XSG-based hardware architecture of gain estimator.

C. Inverse Short-time Fourier Transform Module

The inverse ISTFT module consists of inverse FFT, overlap and add, and normalization subsystem. The inverse FFT is implemented with the Xilinx FFT block with the setting of 512 transform length. However, the two input signals of the inverse FFT block (the real and imaginary parts) have to be converted carefully to the unity scale due to the input normalize requirements. In this paper, the FFT and the inverse FFT are implemented with the 3-multiplier structures for the complex number multiplication for the purpose of resource optimization. Since the overlap-add algorithm utilizes a matrix concatenate block and a submatrix block from the Simulink blocksets, this paper presents simple hardware architectures having two delay chains and a control unit for the matrix concatenate and the submatrix as shown in Fig. 7 and Fig. 8, respectively. By this way, the proposed hardware architecture, which can be expandable for the large buffer size and comprising of only multiple cascaded units of the delay line, can be easily achieved by utilizing simple DSP blocksets of the provided XSG library. The long delay with large data

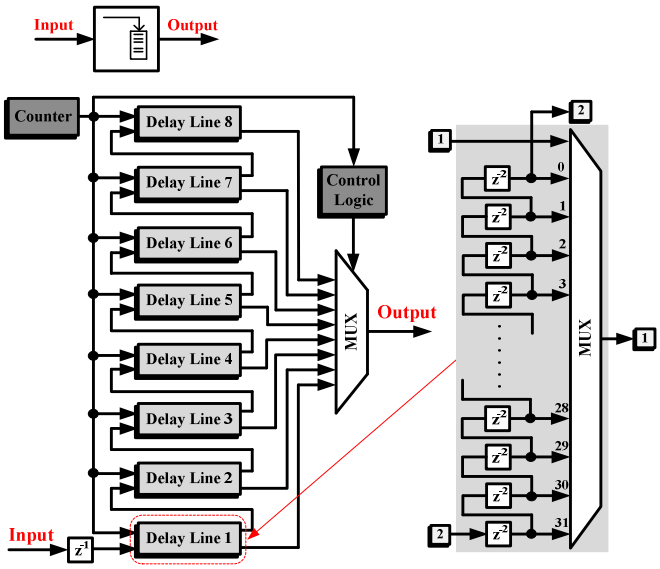

Fig. 5 The simplified architecture of buffer and overlap function.

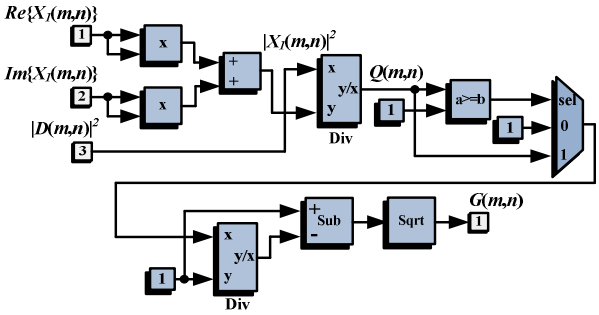

Fig. 6 XSG-based hardware architecture of gain estimator.

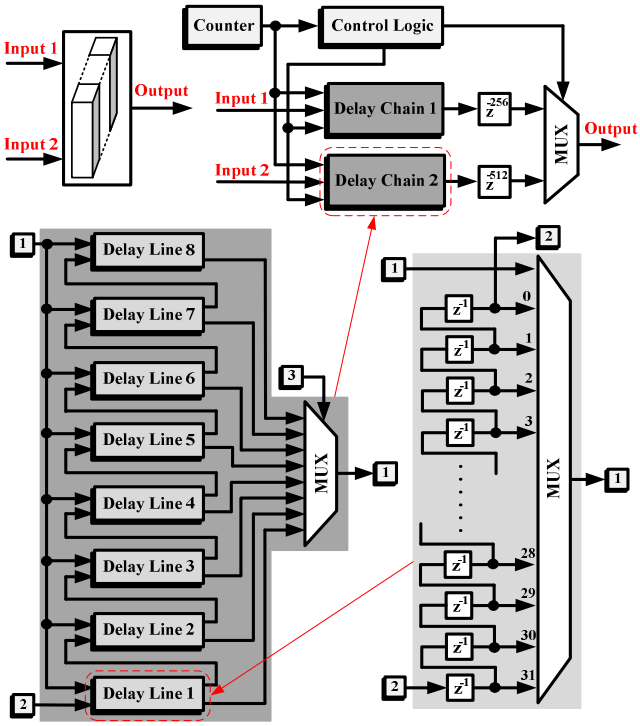

Fig. 7 The simplified architecture of matrix concatenation function.

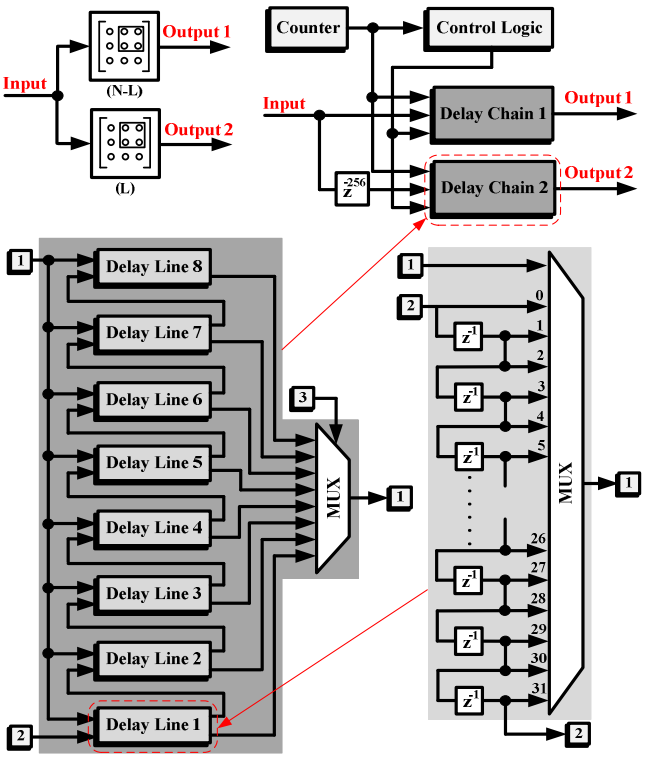

Fig. 8 The simplified architecture of submatrix function.

TABLE I
SIMULATION PARAMETERS OF SHORT-TIME SPECTRAL SUBTRACTION

Input Speech Signal			
Resolutions	16 bits	Sampling Frequency	22.05 kHz
Noise Source			
Noise Type	Gaussian with zero mean	Varian	0.03 @ 0.0 s 0.07 @ 0.5 s 0.05 @ 1.5 s
Gain Estimator			
Control parameter (β)		15	
STFT and Inverse STFT (Hamming window)			
Window length	512	Overlap	256
FFT length	512	Samples per frame	512

bus widths are implemented with a dual port RAM associated with two address counters instead of using a simple delay block provided by the XSG library for minimizing the usage of FPGA hardware resources.

V. SIMULATION RESULTS

In this section, the simulation parameters of the MATLAB-based model and the XSG-based model of dual-channel short-time spectral subtraction in Simulink environment are summarized in the Table I. In order to verify the validity of

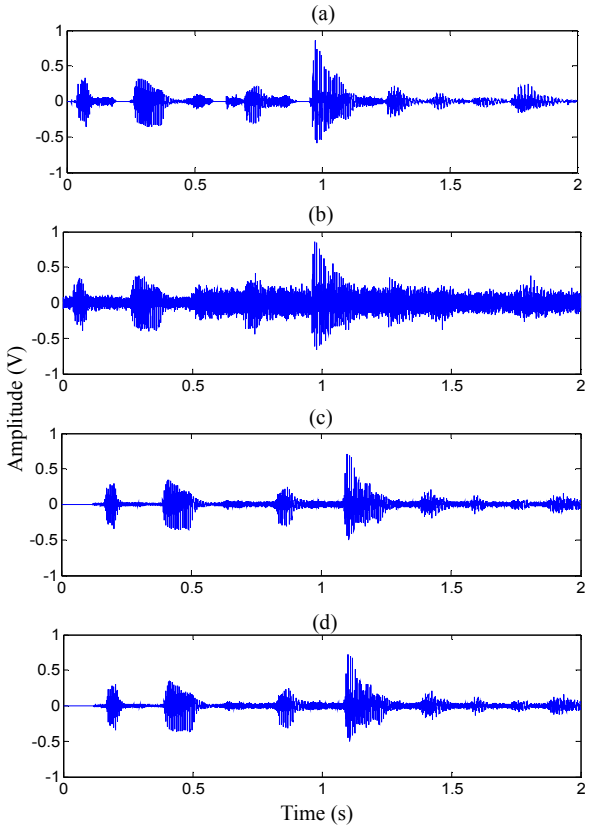

Fig. 9 Simulated speech waveforms: (a) Original speech; (b) Noisy speech; (c) Enhanced speech with MATLAB implementation; (d) Enhanced speech with XSG implementation.

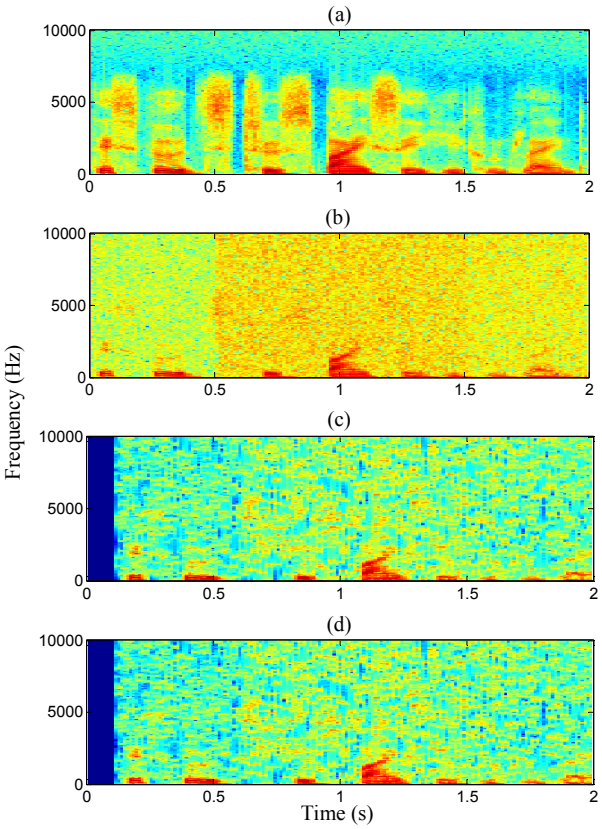

Fig. 10 Spectrograms: (a) Original speech; (b) Noisy speech; (c) Enhanced speech with MATLAB implementation; (d) Enhanced speech with XSG implementation.

TABLE II
FPGA RESOURCE UTILIZATION FOR SHORT-TIME SPECTRAL SUBTRACTION ON XILINX VIRTEX-5 XC5VLX110T CHIP.

Resource Type	Available	Used	Utilization
Number of Flip Flops	69,120	40,703	58%
Number of LUTs	69,120	33,224	48%
Number of Occupied Slices	17,280	13,442	77%
Number of Bonded IOBs	640	39	6%
Number of Block RAM/FIFO	148	20	13%
Number of DSP48Es	64	57	89%
Maximum Operating Freq.	100.806 MHz		

spectral subtraction algorithm for hands-free speech applications, the simulation results are obtained with three different noisy environments starting from the condition (t = 0 s) with the speech signal containing the additive white Gaussian noise (variance of 0.03). After t = 0.5 s, the power spectrum noise is increased with the variance of 0.07. At t = 1.5 s, the noise variance is decreased to 0.05. It should be noted that the additive white noise for each microphone channel is generated separately from different noise sources except the same setting value of the noise variance. It can be observed that the short-time spectral subtraction algorithm has effectively attenuated the acoustic ambient noise and hence significantly enhanced the intelligibility of the target speech. The enhanced speech is adaptively reduced varying additive background noise levels while contains the information very similar to the original speech. In addition, in order to verify the algorithm correctness of the hardware design incorporating with proposed architectures for the FPGA implementation, the simulation results have been compared between the floating-point MATLAB implementation and the fixed-point XSG implementation. It can be concluded that the fixed-point XSG implementation agrees with the floating-point MATLAB implementation and performs consistently the same performance in term of calculation accuracy with the maximum error less than 0.90625 %. Fig 9 and Fig. 10 show the simulated speech waveforms and spectrograms comparing among the original speech, noisy speech, enhanced speech (MATLAB and XSG implementation), respectively. Furthermore, the entire system design of dual-channel short-time spectral subtraction algorithm has been realized and fully integrated into a single target FPGA chip. The fixed-point XSG-based hardware design utilizes approximately with 58% of flip-flops, 48% of LUTs, and 77% of occupied slices available in total of FPGA resources with the maximum operating frequency of 100.806 MHz on the Virtex-5 XC5VLX110T FPGA chip (see Table II), while the design in [9] achieved with the maximum operating frequency of 16.560 MHz on the Virtex-6 XC6VLX240T FPGA platform. The summary of total FPGA resource utilizations confirms that the entire complete system incorporating with proposed architectures is perfectly fitted and possible configured into the target FPGA, which made it is suitable for the rapid prototype development system.

VI. CONCLUSIONS

In this paper, the design of a FPGA-based rapid prototype short-time spectral subtraction algorithm using the XSG tool has been presented. A dual-channel speech enhancement based on the power spectral subtraction for hands-free speech applications was designed in a system-level simulation model using MATLAB Simulink. The XSG-based system model was developed for the FPGA hardware implementation. The hardware architectures for data framing and overlapping algorithms were designed and proposed, which do not provide by the XSG library. The fixed-point algorithm design was verified by comparing results with the floating-point MATLAB implementation, and was successfully realized into a Xilinx Virtex-5 FPGA. This design achievement confirms the implementation feasibility with less design time of the FPGA-based dual channel spectral subtraction algorithm in real-time processing for hands-free speech applications.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[2] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, Jul. 1995.

[3] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Proc. ICASSP*, pp. 493-496, Apr. 1985.

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

[5] S. Guangji, P. Aarabi, and J. Hui, "Phase-Based Dual-Microphone Speech Enhancement Using A Prior Speech Model," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 109-118, Jan. 2007.

[6] B. Handong, Z. Zhiguo, H. Jing, and L. Zhiwen, "FPGA implementation of ICA algorithm for adaptive noise canceling," in *Proc. iCAST*, pp. 452-456, Sep. 2011.

[7] M. Bahoura and H. Ezzaidi, "FPGA-implementation of a sequential adaptive noise canceller using Xilinx System Generator," in *Proc. ICM*, pp. 213-216, Dec. 2009.

[8] M. Bahoura and H. Ezzaidi, "FPGA implementation of a feature extraction technique based on Fourier transform," in *Proc. ICM*, pp. 1-4, Dec. 2012.

[9] M. Bahoura and H. Ezzaidi, "Implementation of spectral subtraction method on FPGA using high-level programming tool," in *Proc. ICM*, pp. 1-4, Dec. 2012.

[10] J. Whittington, K. Deo, T. Kleinschmidt, and M. Mason, "FPGA implementation of spectral subtraction for in-car speech enhancement and recognition," in *Proc. ICSPCS*, pp. 1-8, Dec. 2008.

[11] J. Whittington, K. Deo, T. Kleinschmidt, and M. Mason, "FPGA implementation of spectral subtraction for automotive speech recognition," in *Proc. CIVVS*, pp. 72-79, Mar. 2009.

[12] U. Mahbub, T. Rahman, and A. B. M. H. Rashid, "FPGA implementation of Real Time acoustic noise suppression by Spectral Subtraction using Dynamic Moving Average Method," in *Proc. ISIEA*, pp. 365-370, Oct. 2009.

[13] O. Cappe and J. Laroche, "Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 84-93, Jan. 1995.