Real-Time Enhancement of Electrolaryngeal Speech by Spectral Subtraction

S. Khadar Basha and Prem C. Pandey
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India
Email: basha0509@gmail.com, pcpandey@ee.iitb.ac.in

Abstract— **An electrolarynx, a vibrator held against the neck tissue, is used by laryngectomy patients to provide excitation to the vocal tract as a substitute to that provided by the glottis. The quality and intelligibility of electrolaryngeal speech is generally poor because of the presence of background noise caused by leakage of acoustic energy from the vibrator and vibrator-tissue interface. This noise can be suppressed by pitch-synchronous application of spectral subtraction. The paper presents a real-time implementation of the spectral subtraction for enhancement of electrolaryngeal speech, using a 16-bit fixed-point DSP board. Electrolaryngeal speech is continuously acquired at 12 kHz using codec and DMA into the input buffers. It is processed using 256-point FFT, 3-frame 4-stage cascaded median-based dynamic estimation of noise, spectral subtraction, and IFFT, using two-pitch period window with 50 % overlap. The resynthesized speech is output using DMA and codec.**

Keywords- Electrolaryngeal speech; Electrolarynx; Spectral subtraction; Real-time implementation

I. INTRODUCTION

Electrolarynx, or the external electronic larynx, is used as a verbal communication aid by laryngectomy patients. It is a battery powered hand-held electronic vibrator. Fig. 1 shows a schematic of speech production using this device. Pulses from its diaphragm, held against the neck, get transmitted through the neck tissue to the vocal tract. The spectrum of the vibrations gets shaped by the time-varying resonances of the vocal tract filter, resulting in electrolaryngeal speech. It is generally corrupted by background noise due to leakage of the energy from the vibrator itself and from the interface between the vibrator diaphragm and the neck tissue. It also suffers from low-frequency spectral deficiency due to attenuation of lower harmonics during the transmission through the neck tissue and monotonicity due to constant pitch [1]-[3].

A model of background noise generation is shown in Fig. 2. Speech signal $s(t)$ and the noise through the leakage path $l(t)$ are generated from the convolution of the excitation $e(t)$ with the impulse response of the vocal tract filter $h_v(t)$ and that of the leakage path $h_l(t)$, respectively. Thus the speech signal and leakage noise are highly correlated. Shielding of the vibrator [4] and several speech

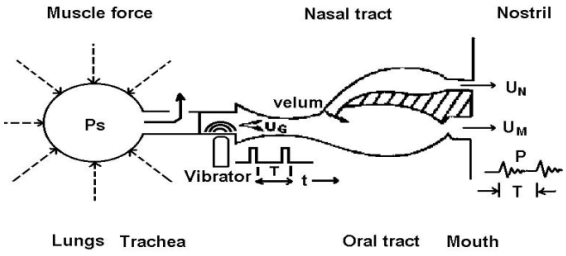

Figure 1. Speech production using an electrolarynx [6].

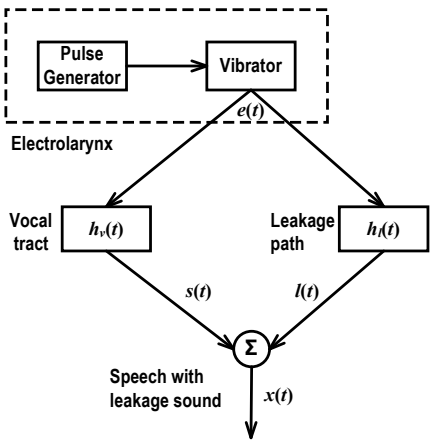

Figure 2. Background noise generation in electro-laryngeal speech [5].

processing techniques [4]-[10] have been used for reducing the background noise. It has been shown in [5] that the pitch-synchronous application of generalized spectral subtraction [11], [12] can be used for suppressing the background noise. In electrolaryngeal speech, identification of non-speech segments using a voicing activity detector is not feasible. Averaging of the spectra during an initial interval of about 2 s with the speaker's lips closed can be used for estimating the leakage noise [5]. But this estimation does not remain valid after a while because change in the pressure and orientation of the device against the neck tissue causes a variation in the spectrum of the leakage noise. For dynamically estimating the noise spectrum, statistical

techniques, like the quantile based noise estimation [13], [14] and minimum statistics based noise estimation [15], [16] can be used. In [17], it has been reported that noise estimation based on the median results in a more consistent speech enhancement than that based on the minimum statistics, and that the naturalness of electrolaryngeal speech can be improved by introducing a small amount of jitter using LPC based analysis-synthesis.

The paper presents an implementation of spectral subtraction using median-based dynamic estimation of magnitude spectrum of the background noise, for real-time enhancement of electrolaryngeal speech, using a 16-bit fixed-point DSP board.

II. SPECTRAL SUBTRACTION

A block diagram of the generalized spectral subtraction is shown in Fig. 3. Windowed frames of the input signal are used for computing the magnitude and the phase spectra. The magnitude spectra of a set of past frames are used to dynamically estimate the magnitude spectrum of the noise. At time index n, input magnitude spectrum $|X_n(k)|$ and the noise magnitude spectrum $L_n(k)$ are used to obtain the enhanced magnitude spectrum $|Y_n(k)|$ by the generalized spectral subtraction

$$E_n(k) = |X_n(k)|^\gamma - \alpha(L_n(k))^\gamma \quad (1)$$

$$\theta_n(k) = \beta\,[L_n(k)]^\gamma \quad (2)$$

$$|Y_n(k)| = [E_n(k)]^{(1/\gamma)}, \text{ if } E_n(k) > \theta_n(k)$$
$$[\theta_n(k)]^{(1/\gamma)}, \text{ otherwise} \quad (3)$$

The resultant magnitude spectrum is combined with the original phase spectrum, and IFFT and overlap-add are used to resynthesize the speech signal. The time-domain signals obtained by IFFT after spectral subtraction generally have discontinuities at frame boundaries. Overlap-add with 50% overlap of analysis windows is used for masking the perceived distortions due to these discontinuities in the resynthesized signal.

The exponent $\gamma = 2$ results in power subtraction and $\gamma = 1$ results in magnitude subtraction. The over-subtraction factor α is used to limit the effect of short-term variations in noise. Over-subtraction may result in negative values of the spectral samples causing 'musical noise', adversely affecting the quality of the output speech. This can be masked by selecting the floor factor β.

It has been shown [5] that use of $\gamma = 1$, i.e. magnitude subtraction, gives the best results for enhancement of electrolaryngeal speech. Optimal values of α and β depend on the method used for noise estimation.

Quantile based noise estimation [6] uses the initial averaging based estimate from an initial silence interval to select the frequency dependent quantiles which are subsequently used to track the noise over a certain number of past frames. This method requires a large number of

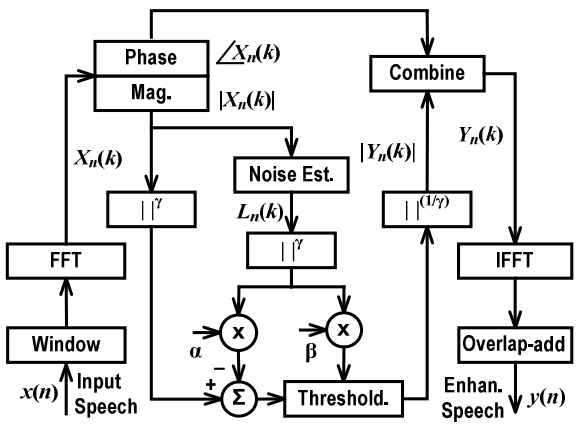

Figure 3. Block diagram of generalized spectral subtraction [17].

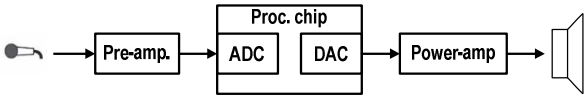

Figure 4. Block diagram of the real-time speech enhancement system.

sorting operations and associated memory and hence it is not suitable for real-time processing using DSP boards. In minimum statistics based noise estimation, the minimum values of the spectral samples over a certain number of past frames are used as the noise estimate [9], [10]. It has a much lower computational and memory requirement. However, in the absence of frequent pauses in speech, speech segments also contribute to the noise estimation and use of a fixed over-subtraction factor results in speech distortion. Median based noise estimation requires relatively more computation, but it estimates the noise dynamically without requiring an initial silence or frequent pauses in speech [17], [18]. For conversational style sentences, the median based dynamic estimation of noise was found to be satisfactory for the number of past frames corresponding to approximately 1.2 s (corresponding to 120 frames for 100 Hz pitch and 2-pitch period windows with 50 % overlap).

III. REAL-TIME IMPLEMENTATION

Figure 4 shows the system for implementing real-time enhancement of electrolaryngeal speech. It uses DSP board "eZdsp" with 16-bit fixed-point processor TMS320C5515 (from Texas Instruments) [19], with on-board stereo codec [20]. The processor has a program memory space of 16 MB and a maximum clock rate of 120 MHz with several options of internal and external clock sources with dynamic switching. Other important on-chip features include 320 KB RAM (including 64 KB dual access RAM), 128 KB ROM, four 4-channel DMA controllers, three 32-bit timers, and FFT hardware accelerator supporting 8 – 1024 point real and complex-valued FFT. The board has TLV320AIC3204

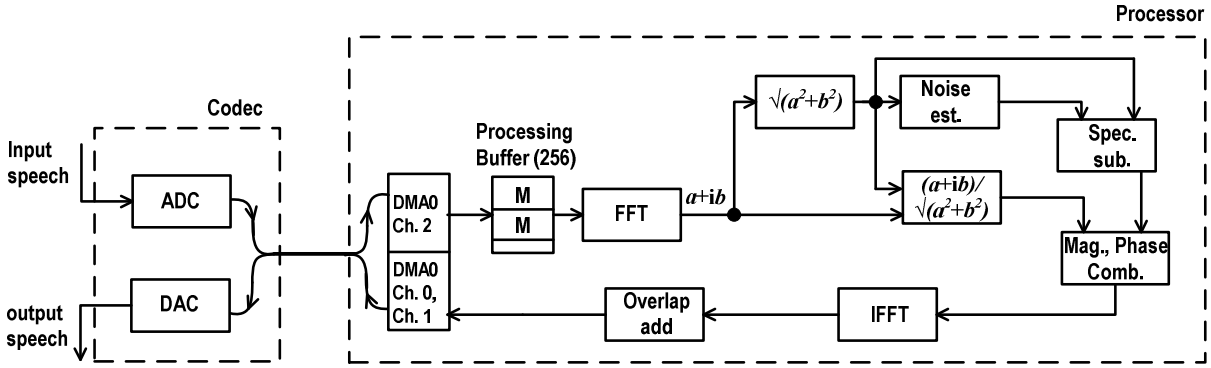

Figure 5. Block diagram of real-time processing on the DSP board.

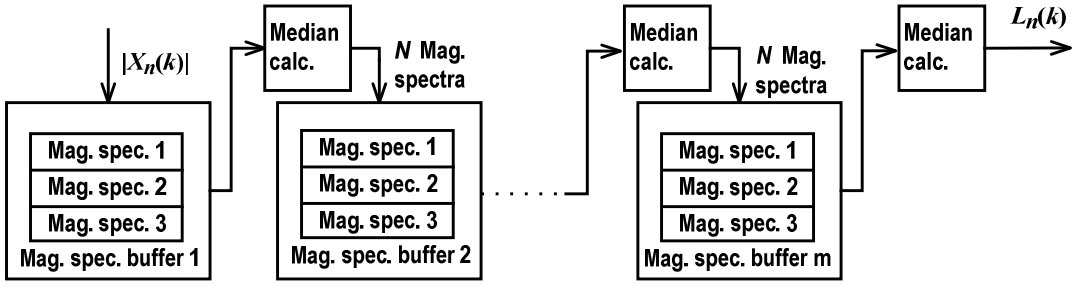

Figure 6. Block diagram of 3-frame m-stage cascaded median of 3^m input frames.

codec with stereo ADC and DAC supporting sampling rate of 8 – 192 kHz. In our implementation, the processor clock is set to 100 MHz and one ADC and one DAC are used at a sampling rate of 12 kHz. Program was written in C, loaded into the program memory of the board, and debugged using the development environment TI 'CCStudio' ver. 4.0.

A block diagram of the implementation is shown in Fig. 5. The input samples are cyclically acquired into two input buffers using DMA for block processing. As one of these buffers gets filled, DMA interrupt occurs and its values are copied to the processing buffer. The frames of the resynthesized speech are alternately placed in two output buffers and DMA is used to cyclically output the samples.

Analysis is carried out using a window length of two pitch periods. Each sample is stored as two pitch periods. As the pitch period of the electrolaryngeal speech remains constant, pitch-synchronous analysis is realized by setting the length M of each of the two input buffers equal to one pitch period. Implementation uses 256-point FFT, permitting pitch down to approximately 90 Hz. An array of 256 words (initialized to zero values) serves as the processing buffer. It is treated as three sub-arrays, first and second of M words each and the third one of 256–2M words. After filling of either of the DMA input buffers, the values in the second sub-array are copied to the first sub-array, and the values in the input buffer are copied to the second sub-array. Thus the analysis window has 2 M samples with 50 % overlap, and these samples get automatically padded with zero-valued samples forming a 256-sample buffer for FFT computation. After FFT, the complex spectral samples are used to calculate the magnitude.

The processing speed and memory constraints of the processor do not permit implementation of 1.2 s running median for noise estimation. To reduce the sorting operations and memory requirement, an n-frame m-stage cascaded median, requiring nm arrays for storing the spectra and giving a result based on n^m frames, was investigated. Figure 6 shows the block diagram for 3-frame m-stage cascaded median approach. Each stage has two-dimensional integer arrays of size 3 × 128. Arrays in the first stage store the magnitude values of the FFTs of the three immediate past frames. After every three frames, an ensemble median is calculated and stored in the array of the next stage. Similar process is applied on each stage till the last stage. In order to limit the maximum computation time taken in each frame, at the most only one median is calculated every frame, giving priority to the higher stages. Therefore the calculation of the medians in a given stage may be missed at certain frames, but this is not likely to affect the estimated noise. It was found that use of 3-frame 4-stage cascaded median results in satisfactory noise estimation.

Earlier investigation [18] using Matlab based implementation showed that for $\gamma = 1$, use of $\alpha = 1.625$ and $\beta = 0$ resulted in good noise reduction without any appreciable musical noise. Hence these values were selected for real-time implementation. The resulting magnitude is combined with the original phase, without explicit phase

calculation, to get the complex spectral value as follows

$$[Y_n(k)]_{real} = |Y_n(k)|\,[X_n(k)]_{real} / |X_n(k)| \qquad (4)$$

$$[Y_n(k)]_{imag} = |Y_n(k)|\,[X_n(k)]_{imag} / |X_n(k)| \qquad (5)$$

Applying overlap-add on IFFT of the resulting complex spectrum gives the output samples. The results are copied alternately to the two output buffers, each of length M samples, for writing to the DAC using DMA.

IV. RESULTS

Figure 7 shows an example of input electrolaryngeal speech and the real-time processed output, as acquired at the input and output of the DSP board. Quality of the processed output from the real-time implementation was found to match that of the Matlab based offline implementation of the spectral subtraction with cascaded median based noise estimation [17], [18]. Processing was effective in suppressing the leakage noise without introducing any significant distortion.

V. CONCLUSION

Spectral subtraction algorithm for enhancement of electrolaryngeal speech has been implemented for real-time operation using a fixed-point DSP board. A 3-frame 4-stage cascaded median is used for dynamic estimation of the noise spectrum. The results have been validated with respect to offline implementation. Use of a faster processor may help in implementing LPC based analysis-synthesis to introduce jitter for increasing the naturalness of the output speech.

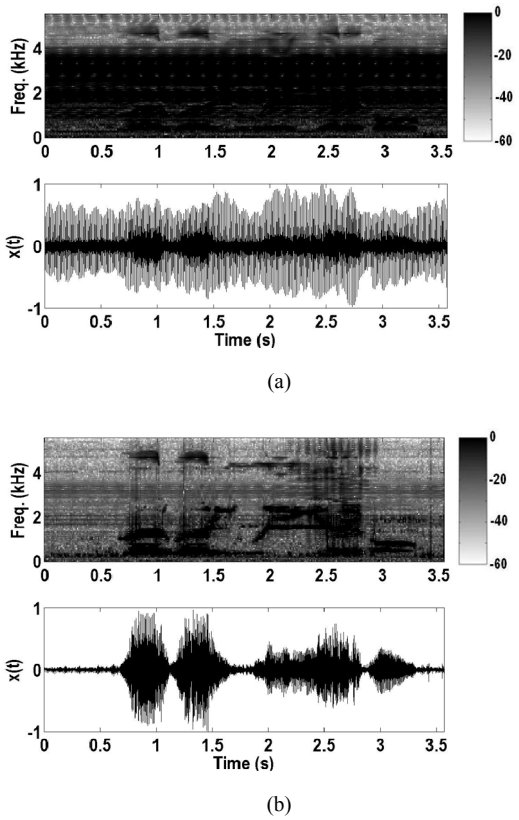

Figure 7. Example of processing: (a) electrolaryngeal speech signal and spectrogram for the sentence ' … *Where were you a year ago ?',* (b) processed output using real-time implementation with 3-frame 4-stage cascaded median based noise estimation, α =1.625, β =0, and γ = 1.

ACKNOWLEDGMENT

The research is supported by the project "National Program on Perception Engineering", sponsored by the Department of Information Technology, MCIT, Government of India.

REFERENCES

[1] H. L. Barney, F. E. Haworth, and H. K. Dunn, "An experimental transistorized artificial larynx," *Bell Systems Tech. J.*, vol. 38, No. 6, pp. 1337-1356, 1959.

[2] M. Weiss, G. Y. Komshian, and J. Heinz, "Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx," *J. Acoust. Soc. Am.*, vol. 65, No. 5, pp. 1298-1308, 1979.

[3] Q. Yingyong and B. Weinberg, "Low-frequency energy deficit in electrolaryngeal speech," *J. Speech and Hearing Research*, vol. 34, pp. 1250- 1256, 1991.

[4] C. Y. Espy-Wilson, V. R. Chari, and C. B. Huang, "Enhancement of alaryngeal speech by adaptive filtering," in *Proc. ICSLP*, 1996, pp. 764-771.

[5] P. C. Pandey, S. M. Bhandarkar, G. K. Baccher, and P. K. Lehena, "Enhancement of alaryngeal speech using spectral subtraction," in *Proc. 14th Int. Conf. Digital Signal Prcessing* (*DSP 2002*), Santorini, Greece, 2002, pp. 591-594.

[6] P. C. Pandey, S. S. Pratapwar, and P. K. Lehana, "Enhancement of electrolaryngeal speech by reducing leakage noise using spectral subtraction with quantile based dynamic estimation of noise", in *Proc. 18th Int. Congress on Acoustics,* (*ICA 2004*), Kyoto, Japan, 2004, 3029-3032.

[7] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Application of spectral subtraction method on enhancement of electrolaryngeal speech," *J.Acoust. Soc. Am.*, vol. 120, pp. 398-406, 2006.

[8] H. Liu, Q. Zhao, M. Wan and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomed. Eng*., vol. 53, pp. 865-874, 2006.

[9] P. Mitra and P.C. Pandey, "Enhancement of electrolaryngeal speech by spectral subtraction with minimum statistics-based noise estimation," (abstract), *J. Acoust. Soc. Amer.*, vol. 120, p. 3039, 2006.

[10] R. Kabir, A. Greenblatt, K. Panetta, and S. Agaian, "Enhancement of alaryngeal speech utilizing spectral subtraction and minimum statistics," in *Proc. 7th Int. Conf. Machine Learning and Cybernetics,* Kunming, China, 12-15 July, 2008.

[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process*., Vol. 27, No. 2, pp. 113-120, 1979.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, 1979, pp. 208-211.

[13] V. Stahl, A. Fisher, and R. Bipus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proc. IEEE ICASSP, 2000*, Vol. 3, pp. 1875-1878.

[14] N. W. D. Evans, J. S. Mason, and B. Fauve, "Efficient real-time noise estimation without speech, non-speech detection: An assessment on the Aurora corpus," in *Proc 14th Int. Conf. Digital Signal Processing* (*DSP 2002*) Santorini, Greece, 2002, pp. 985-988.

[15] R. Martin, "Spectral subtraction based on minimum statistic," in *Proc. 7th European Signal Processing Conf., EUSIPCO-94*, Edinburgh, Scotland, 13-16 September 1994, pp. 1182-1185.

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 5, pp. 504-512, July 2001.

[17] P. C. Pandey and S. K. Basha, "Enhancement of electrolaryngeal speech by spectral subtraction, spectral compensation, and introduction of jitter and shimmer," in *Proc. 20th International Congress on Acoustics* (*ICA 2010*), Sydney, Australia, 2010, Paper no. 670.

[18] S. K. Basha, "Enhancement of electrolaryngeal speech", M.Tech. thesis, Electrical Engineering, Indian Institute of Technology Bombay, 2011.

[19] Texas Instruments Inc., "TMS320C5515 Fixed-Point Digital Signal Processor," 2011, [online] Available: http://focus.ti.com/lit/ds/symlink/tms320c5515.pdf.

[20] Texas Instruments Inc., "TLV320AIC3204 Ultra Low Power Stereo Audio Codec," 2008, [online] Available: http://focus.ti.com/lit/ds/symlink/tlv320aic3204.pdf.