Speech Enhancement Using Spectral Subtraction and Cascaded-Median Based Noise Estimation for Hearing Impaired Listeners

Santosh K. Waddi, Prem C. Pandey, and Nitya Tiwari
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India
Email: {wsantosh, pcpandey, nitya} @ ee.iitb.ac.in

***Abstract*— A spectral subtraction technique is presented for real-time speech enhancement in the aids used by hearing impaired listeners. For reducing computational complexity and memory requirement, it uses a cascaded-median based estimation of the noise spectrum without voice activity detection. The technique is implemented and tested for satisfactory real-time operation, with sampling frequency of 12 kHz, processing using window length of 30 ms with 50% overlap, and noise estimation by 3-frame 4-stage cascaded-median, on a 16-bit fixed-point DSP processor with on-chip FFT hardware. Enhancement of speech with different types of additive stationary and non-stationary noise resulted in SNR advantage of 4 – 13 dB.**

***Keywords*— hearing aid; real-time processing; spectral subtraction; speech enhancement**

I. INTRODUCTION

Hearing aids provide automatic volume control and frequency-selective amplification to compensate for the hearing loss. They may use directional microphones and filtering for reducing environmental noise. Signal processing in digital hearing aids generally involves multi-channel dynamic range compression with selectable channels and settable values of attack time, release time, and compression ratios [1] – [3]. Sensorineural impairment is generally associated with increased spectral masking due to widened auditory filters. Several techniques, such as binaural dichotic presentation [4], [5], spectral contrast enhancement [6], and multiband frequency compression [7], [8], have been reported for reducing the adverse effect of increased spectral masking on speech perception. Despite these advances, hearing aid users with sensorineural impairment experience great difficulty in speech perception in noisy environments. Similar difficulty is faced by users of cochlear prostheses and other sensory aids for the hearing impaired [1], [9]. Use of a second microphone in these aids to provide reference input for noise suppression by adaptive filtering is impractical. Hence, single-input noise suppression is the most practical solution for improving speech quality and intelligibility.

The noise suppression technique should have low algorithmic delay and low computational complexity to permit its implementation on a low-power processor in a sensory aid. Spectral subtraction is a single-input speech enhancement technique developed for use in audio codecs and speech recognition [10]. It involves estimating the noise spectrum, subtracting it from the noisy speech spectrum, and re-synthesizing the speech signal. As the interfering noise is non-stationary, its spectrum needs to be dynamically estimated. Under-estimation of the noise results in residual noise and its over-estimation results in distortion leading to degraded quality and reduced intelligibility. Noise can be estimated during the silence intervals identified by voice activity detection [10]. But the detection may not be satisfactory under low-SNR conditions and the method may not correctly track the noise spectrum during long speech segments. Several statistical techniques for estimating the noise spectrum, without voice activity detection, have been reported [10] – [14]. Their computational complexity and memory requirement pose difficulty in real-time processing using a low-power processor.

A spectral subtraction technique for speech enhancement using cascaded-median based continuous updating of the noise spectrum, without using voice activity detection, is presented. It is implemented for real-time operation on a 16-bit fixed-point DSP processor, with on-chip FFT hardware. The following sections present the signal processing method for noise suppression, its implementation for real-time processing, test results, and conclusions.

II. SIGNAL PROCESSING FOR SPECTRAL SUBTRACTION

Spectral subtraction for enhancement of speech corrupted by additive noise involves estimating the magnitude spectrum of the noise, using it for estimating the magnitude spectrum of the speech signal, and re-synthesizing the speech using the enhanced magnitude spectrum along with the phase spectrum of the noisy speech. Several investigations have been reported, providing different methods for each of these steps [10], [15] – [19]. The noise is estimated during non-speech segments using a voice activity detector or it is dynamically estimated using statistical methods. A block diagram of speech enhancement using spectral subtraction is shown in Fig. 1. It involves windowing, FFT calculation, noise spectrum estimation, spectral subtraction, complex spectrum calculation, and re-synthesis using IFFT with overlap-add.

The magnitude and phase spectra are calculated from FFT of the windowed frames of the noisy speech signal $x(n)$. The magnitude spectra of the past frames are used to estimate the noise magnitude spectrum $D_n(k)$. The enhanced magnitude spectrum $|Y_n(k)|$ is computed, using generalized spectral subtraction, as

$$|Y_n(k)| = \beta^{1/\gamma} D_n(k), \quad \text{if } |X_n(k)| < (\alpha+\beta)^{1/\gamma} D_n(k) \\ [|X_n(k)|^\gamma - \alpha(D_n(k))^\gamma]^{1/\gamma} \quad \text{otherwise} \tag{1}$$

Here γ is an exponent factor, resulting in power subtraction for γ = 2 and magnitude subtraction for γ = 1. Use of subtraction factor $\alpha > 1$ reduces the broadband peaks in the residual noise, but it may result in deep valleys, causing warbling or musical noise and adversely affecting the speech quality. The musical noise is masked by a floor noise controlled by the spectral floor factor β. These two factors offer a great flexibility in the algorithm. Several methods, with different computational complexity, using frequency-dependent factors and factors as functions of *a posteriori* estimate of SNR have been reported [10].

Assuming that the phase error does not significantly affect the intelligibility and quality of speech, the enhanced magnitude spectrum is combined with the original noisy phase, to get the complex spectrum

$$Y_n(k) = |Y_n(k)| e^{j\angle X_n(k)} \tag{2}$$

In order to avoid phase calculation, the complex spectrum is calculated using

$$Y_n(k) = |Y_n(k)| X_n(k) / |X_n(k)| \tag{3}$$

The resulting complex spectra are used to re-synthesize the speech signal. As spectral subtraction involves modification of short-time Fourier transform, there may be discontinuities between signal segments corresponding to the modified complex spectra of the consecutive frames. Use of overlap-add in the re-synthesis helps in masking them. Details of the noise estimation and re-synthesis of enhanced signal are described in the following subsections.

A. Noise estimation

Minimum statistics based noise estimation [11] tracks the noise as minima of the magnitude spectra of the past frames. It is suitable for real-time operation, but it often underestimates the noise and needs estimation of an SNR-dependent subtraction factor. In absence of significant silence segments, it may result in removal of some parts of speech signal during the weaker segments. Quantile-based estimation [14] is based on the observation that the signal energy in a particular frequency bin is low in most of the frames and high only in 10-20% frames corresponding to voiced speech segments. Therefore the noise spectrum samples may be dynamically estimated as a certain quantile value from the histogram of the previous frames. Out of the several frequency and SNR dependent methods for quantile selection, a median-based estimation has been reported to work in a robust manner [14]. As sorting operations require a large amount of computation and memory, the method is not suitable for real-time operation.

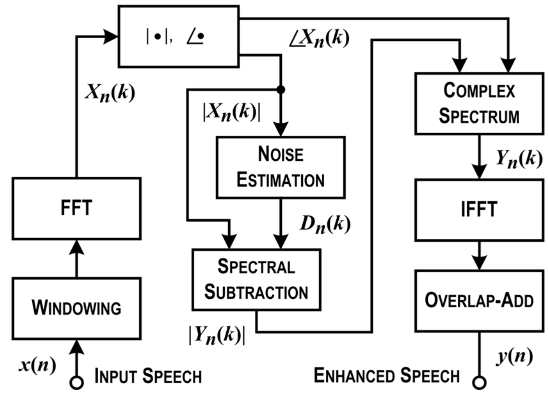


Figure 1. Speech enhancement by spectral subtraction

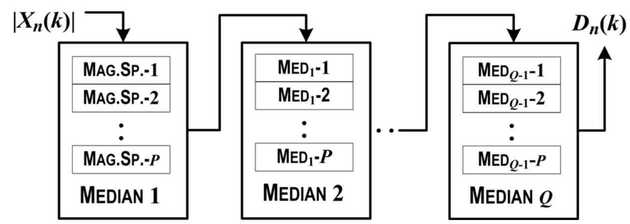


Figure 2. A p-point q-stage cascaded-median based noise estimation

A cascaded-median [18] can be used as an approximation to median, with a significantly reduced computation and memory requirement. In a p-frame q-stage cascaded-median, as shown in Fig. 2, each stage has a first-in-first-out buffer holding p magnitude spectra. The first stage receives the input-frame spectrum. After every p inputs, an ensemble median is calculated and given as input to the next stage. The same process is followed in all the stages and the output of the last stage is taken as an approximation of the ensemble median of the spectra over p^q past frames. Let us compare the number of sorting operations and storage per frequency bin, assuming that the noise spectrum is estimated every M frames from the previous M frames. True-median requires M-sample array for buffering and M-sample array for sorting. For arranging the samples in ascending order, it requires a total of $M(M–1)/2$ sorting operations, i.e. $(M-1)/2$ operations per frame. With $M = p^q$, the cascaded-median requires q p-sample arrays. It results in a storage saving ratio of $2M/(pq)$, and $q \approx \ln(M)$ gives the highest saving. For uniformity in the number of computational operations across frames, median is calculated in only one stage at each frame position, giving priority to the higher stage. In this method, some frames do not contribute to the median calculation, but this fact does not significantly affect the noise estimation. In this case, $p(p-1)/2$ sorting operations are needed per frame. Thus the saving ratio for sorting operation per frame is $(M-1)/p(p-1)$. A lower p results in lesser computation and p = 3 simplifies the programming for sorting operations.

Investigations using typical speech signals and several types of noise (white noise, pink noise, babble, car noise, and train noise) showed that the noise spectrum should be estimated by taking median over frames corresponding to

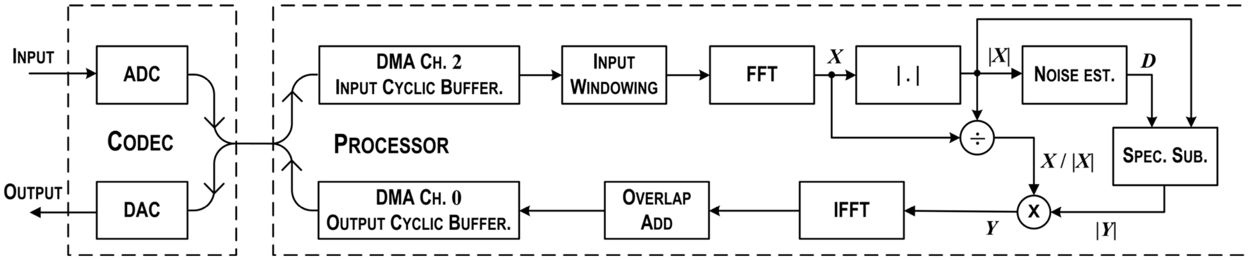


Figure 3. Implementation of spectral subtraction on the DSP board

approximately one second. Use of a 3-frame 4-stage cascaded-median approach (M = 81, p = 3, q = 4) and 30 ms windowing with 50% overlap gives an approximation of median of frames over past 1.215 s. Informal listening showed the results of spectral subtraction obtained using the noise spectrum estimated by the true-median and the cascaded-median approaches to be indistinguishable. For each frequency bin, use of cascaded-median reduced the storage requirement from 162 samples to 12 samples and the number of sorting operations per frame from 40 to 3.

B. Re-synthesis of enhanced signal

The magnitude spectrum of the noisy speech and the estimated magnitude spectrum of the noise are used to get the clean magnitude spectrum using (1). The resulting magnitude spectrum is combined with the original noisy phase spectrum, using (3), to get the complex spectrum which is used for estimating the enhanced speech signal.

To examine the effect of the processing parameters, the technique was implemented using Matlab for offline processing. Implementation using 50% and 75% overlap resulted in similar enhanced output and hence 50% overlap is selected for real-time implementation. It was observed that for different noises and SNR values, appropriate selection of subtraction factor α and floor factor β resulted in almost similar results for magnitude subtraction (exponent factor γ = 1) and power subtraction (γ = 2). The results of magnitude subtraction showed higher tolerances to variation in the values of α and β, and hence only magnitude subtraction was used for real-time implementation. Processing was carried out with sampling frequency of 12 kHz and 30 ms frames (frame length L = 360 samples). As the processed outputs with FFT length N = 512 and higher were indistinguishable, N = 512 has been used in real-time processing. The technique has algorithmic delay of one frame. For real-time processing, the processing delay should not exceed the frame shift. Thus 30 ms frame and 50% overlap-add results in a delay of 45 ms.

III. IMPLEMENTATION FOR REAL-TIME PROCESSING

In order to use it in sensory aids for the hearing impaired [1], [3], the spectral subtraction technique presented in the previous section needs to be implemented for real-time operation on a low-power DSP chip. The technique is implemented on the 16-bit fixed point processor TI/TMS320C5515 [20]. The processor has 16 MB memory space with 320 KB on-chip RAM (including

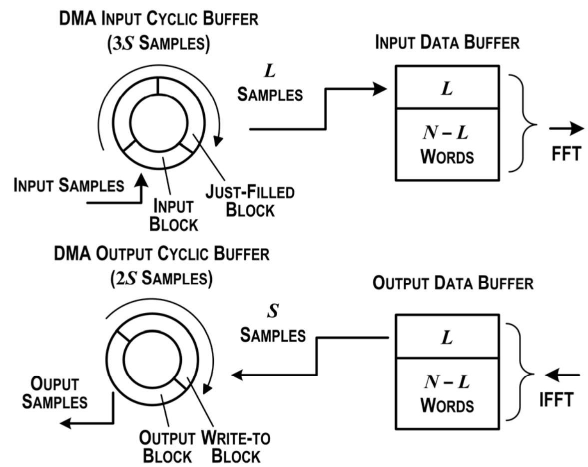


Figure 4. Data transfer and buffering operations ($S = L/2$).

64 KB dual access RAM), 128 KB on-chip ROM. It has three 32-bit programmable timers and four DMA controllers each with four channels. It has a tightly coupled hardware accelerator supporting 8 to 1024-point FFT. The maximum clock rate is 120 MHz. The implementation was carried out using DSP board "eZdsp" [21], with 4 MB on-board NOR flash for user program and codec TLV320AIC3204 [22] with stereo ADC and DAC supporting 16/20/24/32-bit quantization and sampling frequency of 8 – 192 kHz. The program was written in C, using TI's 'CCStudio, ver. 4.0' as the development environment. The implementation uses one channel of the codec, with 16-bit quantization and 12 kHz sampling.

A block diagram of the implementation, with L-sample window and N-point FFT (L = 360, N = 512), is shown in Fig. 3. At the set sampling frequency, DMA channel-2 reads the ADC values into the input cyclic buffer and channel-0 writes the output cyclic buffer values to DAC. The input samples, spectral values, and the processed samples are all stored as 4-byte words with 16-bit real and 16-bit imaginary parts, in order to reduce the conversion overheads. The input, output, data transfer, and buffering operations are devised for an efficient realization of the processing with 50% overlap and zero padding. As shown in Fig. 4, the input samples are acquired using a 3-block input cyclic buffer and the processed samples are output using a 2-block cyclic buffer, with S-word blocks and $S = L/2$. Pointers with cyclic values (.., 1, 2, 3, 1, ..) are used to track the current input and just-filled input blocks. They are initialized to 1 and 3, respectively. The current output and write-to output blocks are tracked by pointers with toggling values of 1 and 2, and initialized as 1 and 2, respectively. A DMA interrupt is generated when the current input block gets filled. All pointers are incremented cyclically. The DMA-mediated reading from ADC and writing to DAC are continued. The samples of the just-filled and the previous blocks are copied to the input data buffer, and are padded with $N-L$ zero-valued samples to serve as input to N-point FFT. The processing for noise estimation, spectral subtraction, and re-synthesis of output signal is implemented with due care to avoid overflows.

IV. TEST RESULTS

The evaluation of the proposed technique and its real-time implementation was carried out using informal listening and objective evaluation using perceptual evaluation of speech quality (PESQ) measure [10], [23]. This objective measure is a prediction of the subjective mean opinion score (MOS) of the degraded speech and is calculated from the difference between the loudness spectra of level-equalized and time aligned original and degraded signals. The speech material consisted of a recording with three isolated vowels, a Hindi sentence, and an English sentence (-/a/-/i/-/u/– "aayiye aap kaa naam kyaa hai?" – "Where were you a year ago?") from a male speaker. For informal listening test, a longer test sequence was generated by speech-speech-silence-speech concatenation of the recording. Testing involved processing of speech with additive white, pink, babble, car, and train noises at SNR of 15, 12, 9, 6, 3, 0, -3, and -6 dB. Informal listening test showed that the processing significantly enhanced the speech for all noises and there was no audible roughness. While spectral floor factor $\beta = 0.001$ was found to be appropriate in all cases, most appropriate value of subtraction factor α varied over 1.5 – 2.5.

An objective evaluation of the outputs from Matlab-based implementation was carried out using PESQ measure for different types of noise and SNR conditions. Fig. 5 shows the PESQ score vs. SNR plot of unprocessed and processed signals for white and babble noises. For un-processed speech, the score decreased progressively with decrease in SNR. The scores for white noise were lower than the corresponding ones for babble noise. While processing of noise-free speech decreased the score from 4.5 to 3.7, processing of noisy speech increased the scores by 0.48 – 0.90 for white noise and by 0.13 – 0.33 for babble noise. For a score of 2.5 (generally considered as lowest score for acceptable speech), processing resulted in SNR advantage of approximately 13 dB for white noise and 4 dB for babble noise. SNR advantage for other types of noise was between these two values. For speech added with different types of noise at 0 dB SNR, Table I gives the scores for the unprocessed and processed speech and the optimal values of α. The processing improved the scores in all cases, resulting in scores close to or higher than 2.5.

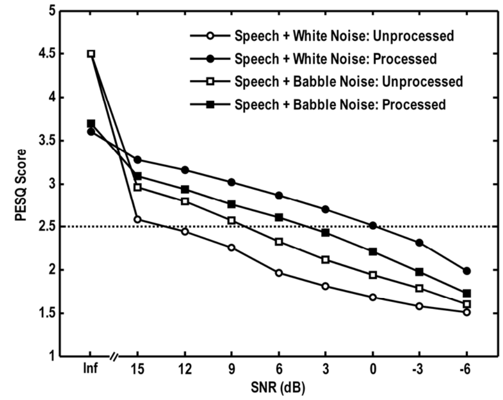


Figure 5. PESQ score vs SNR for noisy and enhanced speech.

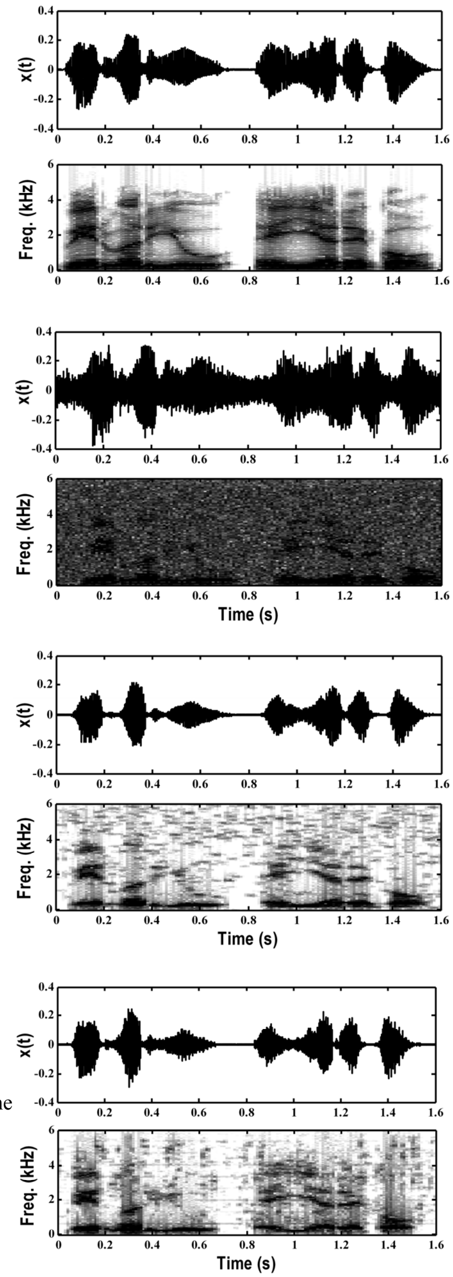
TABLE I. PESQ SCORE OF THE UNPROCESSED AND PROCESSED SPEECH SIGNALS, SNR = 0 DB.

Noise	PESQ Score		Optimal α
	Un-processed	Processed	
White	1.69	2.55	2.5
Pink	1.79	2.60	2.5
Babble	2.13	2.50	2.0
Car	1.77	2.41	2.5
Train	2.08	2.65	2.5

The real-time processing was tested using speech mixed with white, babble, car, pink, and train noises at different SNRs [24]. The speech signal with added noise was output from the PC sound card as input to the codec of the DSP board and its output was acquired through the PC sound card. An example of processing, showing the noise-free speech, noisy speech with white noise at 3 dB SNR, output from offline processing, and output from real-time processing, is shown in Fig. 6. The enhanced speech outputs from the two types of processing were found to be perceptually similar. A comparison of input and output using a DSO showed a processing delay of 48 ms (45 ms corresponding to 1.5 frame and 3 ms due to DMA mediated I/O). The processor has maximum clock frequency of 120 MHz and the speech enhancement was found to be satisfactory for a clock frequency down to 16.4 MHz, indicating that the technique needed approximately 14% of the processing capacity at the clock frequency of 120 MHz and the rest could be used in implementing other processing as needed for a sensory aid.

V. CONCLUSION

A spectral subtraction technique for suppression of additive noise, using cascaded-median based noise estimation for reducing computational complexity and memory requirement, has been presented. Enhancement of speech with different types of additive stationary and non-stationary noise resulted in SNR advantage of 4 – 13 dB. The technique has been implemented on the 16-bit fixed-point processor TI/TMS320C5515, using about one-seventh of its processing capacity. Use of subtraction and spectral floor factors dependent on frequency and *a posteriori* SNR estimate may improve its performance. The proposed speech enhancement technique may be combined



(a) Clean speech signal

(b) Noisy speech signal

(c) Offline processing output

(d) Real-time processing output

Figure 6. Processing of "Where were you a year ago", from a male speaker, with white noise at 3 dB SNR: signals and spectrograms.

with other signal processing techniques used in the sensory aids and tested for improving perception of different speech materials by the hearing-impaired listeners. The implementation using other processors may also be investigated.

ACKNOWLEDGMENT

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

REFERENCES

[1] H. Levitt, J. M. Pickett, and R. A. Houde (eds.), *Senosry Aids for the Hearing Impaired.* New York: IEEE Press, 1980.

[2] J. M. Pickett, *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology*. Boston, Mass.: Allyn Bacon, 1999, pp. 289–323.

[3] H. Dillon, *Hearing Aids*. New York: Thieme Medical, 2001.

[4] T. Lunner, S. Arlinger, and J. Hellgren, "8-channel digital filter bank for hearing aid use: preliminary results in monaural, diotic, and dichotic modes," *Scand. Audiol. Suppl*., vol. 38, pp. 75–81, 1993.

[5] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, "Binaural dichotic presentation to reduce the effects of spectral masking in moderate bilateral sensorineural hearing loss," *Int. J. Audiol*., vol. 51, no. 4, pp. 334–344, 2012.

[6] J. Yang, F. Luo, and A. Nehorai, "Spectral contrast enhancement: Algorithms and comparisons," *Speech Commun*., vol. 39, no. 1–2, pp. 33–46, 2003.

[7] T. Arai, K. Yasu, and N. Hodoshima, "Effective speech processing for various impaired listeners," in *Proc. 18th Int. Cong. Acoust*., 2004, Kyoto, Japan, pp. 1389–1392.

[8] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, "Multi-band frequency compression for improving speech perception by listeners with moderate sensorineural hearing loss," *Speech Commun.*, vol. 54, no. 3 pp. 341–350, 2012.

[9] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," in A. R. Moller (ed.), *Cochlear and Brainstem Implants*, Adv. Otorhinolaryngol. vol. 64, Basel: Karger, 2006, pp. 109–143.

[10] P. C. Loizou, *Speech Enhancement: Theory and Practice*. New York: CRC, 2007.

[11] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Process. Conf.*, 1994, pp. 1182-1185.

[12] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process*., vol. 11, no. 5, pp. 466-475, 2003.

[13] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE ICASSP*, 1995, pp. 153-156.

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

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

[16] 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.

[17] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453-466, 2008.

[18] S. K. Basha and P. C. Pandey, "Real-time enhancement of electrolaryngeal speech by spectral subtraction," in *Proc. Nat. Conf. Commun. 2012*, Kharagpur, India, pp. 516-520.

[19] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, no. 5, pp. 450–475, 2010.

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

[21] Spectrum Digital, Inc., "TMS320C5515 eZdsp USB Stick Technical Reference," 2010, [online] Available: support.spectrumdigital.com/boards/usbstk5515/reva/files/usbstk5515_TechRef_RevA.pdf

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

[23] ITU, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Rec*., P.862, 2001.

[24] S. K. Waddi, "Speech enhancement results", 2013, [online] Available: www.ee.iitb.ac.in/~spilab/material/santosh/ncc2013.