7th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2003), Orlando, USA, July 27-30, 2003

Reduction of Background Noise in Alaryngeal Speech using Spectral Subtraction with Quantile Based Noise Estimation

Santosh S. Pratapwar, Prem C. Pandey, and Parveen K. Lehana
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai-400 076, India
E-mail: {santosh, pcpandey, lehana}@ee.iitb.ac.in

ABSTRACT

The transcervical electrolarynx or external electronic larynx is used by persons who cannot use their natural voice production mechanism. The device is a vibrator held against the neck tissue, and the vibrations generated move up the vocal tract to produce alaryngeal speech. Presence of background noise, due to leakage of the acoustic energy from the vibrator, degrades the resulting alaryngeal speech. This paper presents a single input technique for reducing the background noise in alaryngeal speech signal, using spectral subtraction in a pitch synchronous manner. Updating of the noise magnitude spectrum is carried out using quantile based noise estimation, which does not require speech/non-speech detection.

Keywords: Artificial Larynx, Alaryngeal Speech Enhancement, Spectral Subtraction, Quantile Based Noise Estimation.

1. INTRODUCTION

A schematic of the normal speech production system is shown in Fig. 1. The air stream from the lungs acts as the energy source. Vocal tract provides spectral shaping of the excitation waveform generated by pulsatile flow caused by the vibration of the vocal cords in the larynx or a turbulent flow caused by a constriction in the oral cavity [1]. In some cases of disease and injury, the larynx is surgically removed by an operation known as laryngectomy, and the patient needs external aids to communicate. An artificial larynx [2],[3] is a device used to provide excitation to the vocal tract, as a substitute to that provided by a natural larynx.

External electronic larynx or transcervical electrolarynx, consisting of an electronic vibration generator, is the widely used type of artificial larynx. It is hand held and pressed against the neck tissue. A schematic of speech production using this device is shown in Fig.2. Vibrations generated get coupled to the neck tissue move up the vocal tract, and get spectrally shaped to produce speech. The device is easy to use and portable. However the speaker needs to control the pitch and volume switches to prevent monotonic speech, and this needs practice. The speech produced is generally deficient in low frequency energy due to low coupling efficiency through the throat tissue [4]. The unvoiced segments generally get substituted by voiced ones. In addition to these, a major problem with this device is that the speech output has a background noise, which degrades the quality of the output speech considerably [5],[6],[7],[8].

Here we present a method for enhancing the alaryngeal speech, with effective reduction of the background noise using quantile based noise estimation and spectral subtraction.

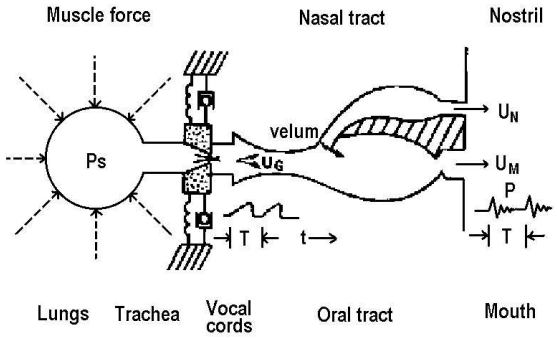


Fig. 1. Schematic of normal speech production [1].

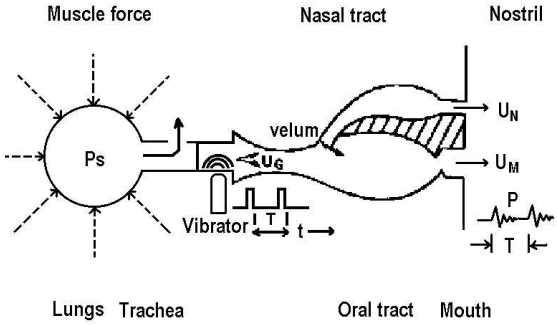


Fig. 2. Schematic of speech production with a transcervical electrolarynx.

2. ENHANCEMENT OF ALARYNGEAL SPEECH

A transcervical electrolarynx generally uses an electromagnetic transducer [6]. The steady background noise is generated due to leakage of the vibrations produced by the vibrator membrane/plate. Front end of the vibrator is coupled to the neck tissue, while the back end is coupled to the air in the instrument housing. Noise is produced due to leakage of acoustical energy from the housing to the air outside. It is present even if the speaker's lips are closed. Vibrations leaked from the front end due to improper coupling of the vibrator to the neck tissue also contribute to the background noise.

Weiss *et al.* [7] reported a detailed study of perceptual and the acoustical characteristics of alaryngeal speech, using the device Western Electric Model 5. Speech-to-noise ratio (SNR), defined as the ratio of the average level of the vocal peaks in the alaryngeal speech (inclusive of background interference) and the level of radiated sound measured with the speaker' s mouth closed, varied over 4-15 dB across the subjects for the same

device. As the leakage from the casing of the instrument should be speaker independent, large variation in SNR indicated that leakage from the vibrator-tissue interface varied across speakers, and it significantly contributed to the background interference. In an early report of the development model of a similar device, Barney *et al*. [6] reported the speech-to-noise ratio to be approximately 20-25 dB. The results of the identification tests revealed that SNR lower than 4 dB had significantly lower intelligibility compared to the speech with higher SNRs. Spectrum of the direct-radiated noise indicated that most of the energy was concentrated in the frequency region 400-800 Hz. A second peak was found between 1-2 kHz, with magnitude down the previous value by 5-10 dB. There were usually 2 or 3 additional peaks between 2 and 4 kHz. The frequency and magnitude of these peaks were speaker dependent. In alaryngeal speech with poor SNRs, there was significant auditory masking of the vowel formants, which could lead to vowel identification errors. However, the noise spectrum was steady in nature in contrast to the rapidly changing formant frequencies. Because of this reason, the listeners were able to track the formant trajectories and perceive speech in the presence of background noise for relatively higher SNRs. However, the background noise reduced the identification of consonants.

Acoustic shielding of the vibrator results only in a marginal reduction of the noise [5]. By using vibrators based on piezoelectric or magnetostrictive effect, the noise can be reduced at the source. However, such vibrators have poor efficiency. Further, vibrator design cannot help in reducing the leakage from the vibrator to tissue interface.

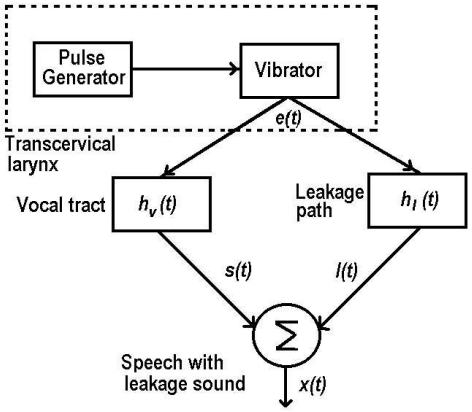


Fig. 3. Model of background noise generation in transcervical electrolarynx [8].

A model of the leakage sound generation during the use of transcervical larynx is shown in Fig.3. The vibrations generated by the vibrator diaphragm have two paths. The first path is through the neck tissue and the vocal tract. Its impulse response $h_v(t)$ depends on the length and configuration of the vocal tract, the place of coupling of the vibrator, the amount of coupling, etc. Excitation $e(t)$ passing through this path results in speech signal $s(t)$. The second path of the vibrations is through the surroundings, and this leakage component $l(t)$ gets added to the useful speech $s(t)$, and deteriorates its intelligibility.

Signal processing techniques can be implemented for reduction of noise by estimating noise present in the signal and subtracting it from the noisy signal. The main problem in noise subtraction is that the speech and noise, resulting from the same excitation, as shown in Fig.3, are highly correlated.

Epsy-Wilson *et al.* [5] reported a technique for enhancement of alaryngeal speech using two-input LMS algorithm. If noise adaptation is carried out during vocal sounds, noise cancellation will result in an output that contains no information at all. During consonantal segments, the correlation between speech and noise gets weaker on account of the vocal excitation being caused by the turbulence at constrictions. Authors have reported that by carrying out noise adaptation during non-sonorant or low energy segments, the noise cancellation was effective, and most of background noise was cancelled. During the sonorant sounds, there was an improvement in the output quality, though the background noise was not removed fully. Processing resulted in improvement in speech intelligibility [5].

A single input noise cancellation technique based on spectral subtraction applied in a pitch synchronous manner has been reported by our group [8]. In this technique, the noise spectrum is estimated by averaging the noise spectra over several segments of the background noise acquired with speaker keeping the lips closed. Because of variations in the noise characteristics, effective cancellation requires frequent acquisitions of noise. We have investigated the use of quantile based noise spectrum estimation for continuous updating of noise spectrum. After a review of the spectral subtraction technique for reducing the background noise [8], use of quantile based noise estimation is presented. This is followed by test results.

3. SPECTRAL SUBTRACTION FOR REDUCING BACKGROUND NOISE

Spectral subtraction technique is one of the important techniques for enhancing noisy speech [9],[10]. The basic assumption in this technique is that the clean speech and the noise are uncorrelated, and therefore the power spectrum of the noisy speech equals the sum of power spectra of noise and clean speech. In case of alaryngeal speech, speech signal and background interference are not uncorrelated. With reference to Fig.3, the noisy speech signal is given as

$$x(n) = s(n) + l(n) \tag{1}$$

where $s(n)$ is the speech signal and $l(n)$ is the background interference or the leakage noise. If $h_v(n)$ and $h_l(n)$ are the impulse responses of the vocal tract path and the leakage path respectively, then

$$s(n) = e(n) * h_v(n) \tag{2}$$

$$l(n) = e(n) * h_l(n) \tag{3}$$

where $e(n)$ is the excitation. Taking short-time Fourier transform on either side of (1), we get

$$X_n(e^{j\omega}) = E_n(e^{j\omega})[H_{vn}(e^{j\omega}) + H_{ln}(e^{j\omega})]$$

Considering the impulse response of the vocal tract and leakage path to be uncorrelated, we get

$$|X_n(e^{j\omega})|^2 = |E_n(e^{j\omega})|^2[|H_{vn}(e^{j\omega})|^2 + |H_{ln}(e^{j\omega})|^2] \tag{4}$$

If the short-time spectra are evaluated using pitch synchronous window, $|E_n(e^{j\omega})|^2$ can be considered as constant $|E(e^{j\omega})|^2$. During non-speech interval, $e(n) * h_v(n)$ will be negligible and the noise spectrum is given as

$$|X_n(e^{j\omega})|^2 = |L_n(e^{j\omega})|^2 = |E(e^{j\omega})|^2 |H_{ln}(e^{j\omega})|^2 \quad (5)$$

By averaging $|L_n(e^{j\omega})|^2$ during the non-speech duration, we can obtain the mean squared spectrum of the noise $|L(e^{j\omega})|^2$. This estimation of the noise spectra can be used for spectral subtraction during the noisy speech segments.

For implementation of the technique, squared magnitudes of the FFT of a number of adjacent windowed segments in non-speech segment are averaged to get the mean squared noise spectrum. During speech, the noisy speech is windowed by the same window as in earlier mode, and its magnitude and phase spectra are obtained. The phase spectrum is retained for resynthesis. From the squared magnitude spectrum, the mean squared spectrum of noise, determined during the noise estimation mode, is subtracted

$$|Y_n(k)|^2 = |X_n(k)|^2 - |L(k)|^2 \quad (6)$$

The resulting magnitude spectrum is combined with the earlier phase spectrum, and its inverse FFT is taken as the clean speech signal $y(n)$ during the window duration

$$y_n(m) = IFFT[|Y_n(k)| \, e^{j\angle Xn(k)}] \quad (7)$$

In real practice, assumptions regarding $h_v(n)$ and $h_l(n)$ being uncorrelated may be valid over long period, but not necessarily over short segments. This may result in some of the frequency components becoming negative, causing narrow random spikes of value between zero and maximum during non-speech segment, known as residual noise. When converted back to the time domain, the residual noise sounds as sum of tone generators with random frequencies turned on and off. During speech period, this noise residual will be perceived at frequencies, which are not masked by the speech. In order to reduce the effect of residual noise, modified spectral subtraction method [10] to reduce spectral excursions is used.

$$\begin{aligned} |Y_n(k)|^2 &= |X_n(k)|^2 - \alpha|L(k)|^2 \\ |Y'_n(k)|^2 &= |Y_n(k)|^2 \quad \text{if } |Y_n(k)|^2 > \beta|L(k)|^2 \\ &= \beta|L(k)|^2 \text{ otherwise} \end{aligned} \quad (8)$$

where α is the subtraction factor and β is the spectral floor factor. With $\alpha > 1$, the noise will be over subtracted from the noisy speech spectrum. This will not only reduce the noise floor, but will also eliminate the peaks of wideband noise, thereby reducing it considerably. However, over-subtraction may lead to the enhancement of the valleys in the vicinity of the peaks, thereby increasing the noise excursion. This is taken care by the spectral floor factor β. The spectral components of $|Y'_n(k)|^2$ are prevented from going below $\beta|L(k)|^2$. For $\beta > 0$, the spectral excursions are not as large as with $\beta = 0$, since the valleys between the peaks are not very deep. This reduces the residual noise to a large extent. The proper choice of the parameters α and β gives an output free from broadband as well as the residual noise.

Another modification by Berouti *et al*. [10] to the spectral subtraction algorithm is the addition of exponent factor γ in place of 2 for subtraction.

$$\begin{aligned} |Y_n(k)|^\gamma &= |X_n(k)|^\gamma - \alpha|L(k)|^\gamma \\ |Y'_n(k)|^\gamma &= |Y_n(k)|^\gamma \quad \text{if } |Y_n(k)|^\gamma > \beta|L(k)|^\gamma \\ &= \beta|L(k)|^\gamma \text{ otherwise} \end{aligned} \quad (9)$$

With $\gamma < 1$, the subtraction of the noise spectrum affects the noisy speech spectrum drastically than with the case when $\gamma = 1$. For $\gamma < 1$, the processed output has a low level, and hence there is a need for normalization of the output level to make it independent of γ[10].

A schematic of the modified spectral subtraction algorithm is shown in Fig.4. Optimal values for reduction of background noise as reported by Pandey *et al.* [8] are: window length= twice the pitch period, spectral subtraction factor $\alpha = 2$, spectral floor factor $\beta = 0.001$, exponent factor $\gamma = 1$.

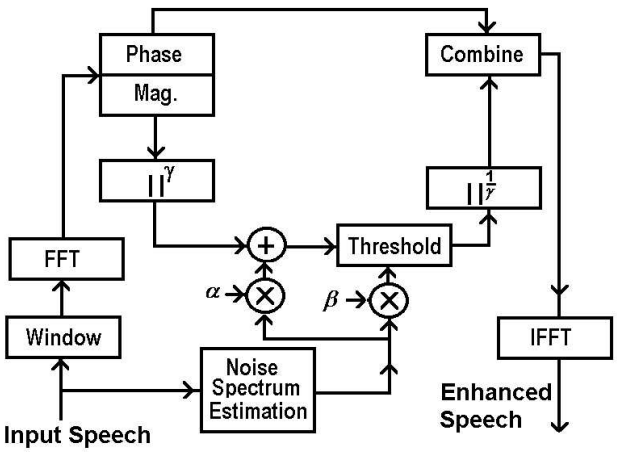


Fig. 4. Modified spectral subtraction algorithm [8].

A speech processor based on this technique will have a noise estimation mode during which speaker keeps the lips closed and the acquired signal consists of only the background noise. After this the device automatically switches to speech enhancement mode: the earlier estimated noise spectrum is used in noise subtraction. The noise spectrum is taken to be constant over the entire duration of enhancement mode. But actually the back-ground noise varies because of variations in the place of coupling of vibrator to the neck tissue and the amount of coupling. This results in variations in the effectiveness of noise enhancement over extended period. Hence a continuous up-dating of the estimated noise spectrum is required. Recursive averaging of spectra during silence segments may be used for noise spectrum estimation [9],[10]. However, speech/silence detection in alaryngeal speech is rather difficult. Quantile-based noise estimation (QBNE) technique [11] does not need speech/non-speech classification and we have investigated its use for noise estimation in alaryngeal speech.

4. QUANTILE BASED NOISE ESTIMATION TECHNIQUE

Quantile-based noise estimation (QBNE) makes use of the fact that even during speech periods, frequency bins tend not to be permanently occupied by speech i.e. exhibit high energy levels [11],[12],[13]. Speech/non-speech boundaries are detected implicitly on a per-frequency bin basis, and noise spectrum estimates are updated throughout speech/non-speech periods. QBNE is simple to implement with relatively few parameters to optimize.

The degraded signal may be analyzed on a frame-by-frame basis, to obtain an array of the magnitude spectral values for each frequency sample, for a certain number of the past frames. Sorting of magnitude values in this array may be used for obtaining a particular quantile value. The rate at which QBNE

reacts to changes in the noise depends on the number of past frames used. If number is too small, the estimation will not be accurate. If the number is too large, reaction to changes will be slow. In this approach, the buffer for all the frequency samples have to be reconstructed and resorted at each frame and this is computationally expensive. For a faster processing, an efficient indexing algorithm [14] was implemented. For each frequency sample w_k, two buffers are used. Sorted value buffer $D(w_k)$ holds the spectral magnitude values for M frames. Index buffer $I(w_k)$ holds the frame number of the corresponding value in $D(w_k)$. After computation of magnitude spectrum of sample values in a new frame, the index of oldest value in $D(w_k)$ is located from $I(w_k)$ and replaced with new value. Frame numbers in $I(w_k)$ are updated. Now there is only one value in $D(w_k)$ that is not in numerical order. It is sorted and $I(w_k)$ is also reordered accordingly. After updating the two buffers, we can find a particular quantile value in $D(w_k)$ for each frequency sample; and these values can be used for quantile based noise estimation continuously.

5. TEST RESULTS

Quantile based noise estimation technique was used for speech enhancement of alaryngeal speech digitally recorded with 16-bit quantization and 11.025 kSa/s sampling rate. Electrolarynx NP-1 (manufactured by NP Voice, India) was used for this purpose. The vibrator of the electrolarynx had a fixed pitch of 90.3 Hz, i.e. pitch period of 122 samples. The degraded signal was analyzed on frame-by-frame basis, with frame size of twice pitch period i.e. 22.1 ms, with 50 % overlap.

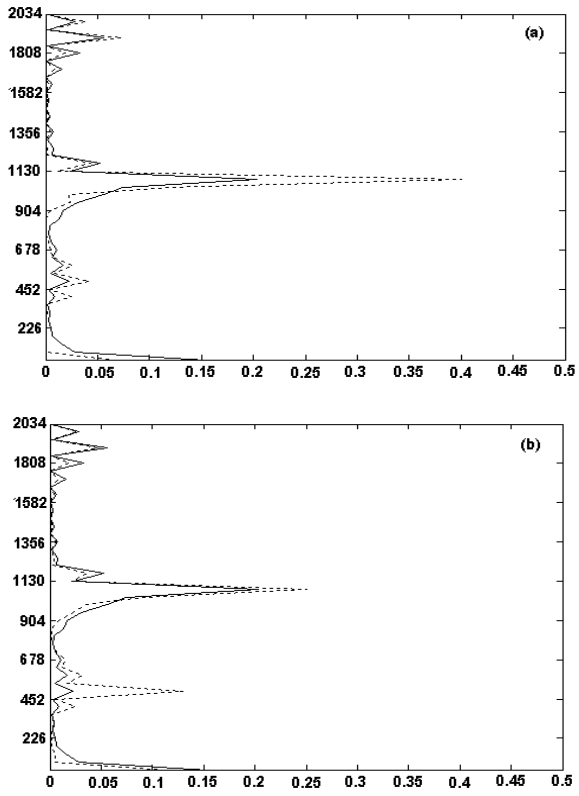
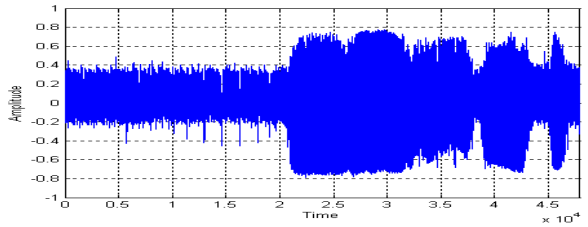


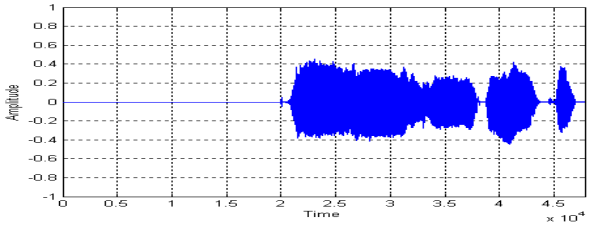
Fig. 5. Plot of average magnitude spectrum of noise and QBNE derived spectrum of noisy speech at (a) 5 percentile (b) 30 percentile. Plot with solid line shows the average magnitude spectrum of noise and the plot with dotted line shows the quantile derived spectrum of noisy speech.

Analysis was carried out to obtain optimum QBNE derived spectrum of noisy speech for effective noise cancellation by comparing different quantile spectra of noisy speech with the average magnitude spectrum of noise only. Averaged noise spectrum was obtained over the initial silence duration of 1.8 s. Quantile spectra of noisy speech were obtained using 55 frames with 50% overlap, i.e. segment length of 0.6 s contributed to updating.

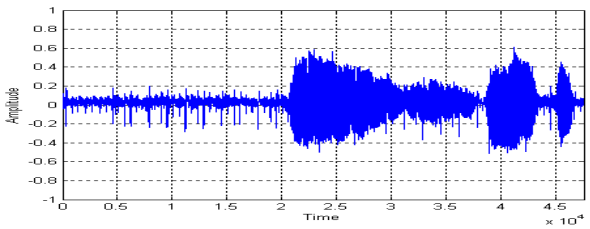
Fig. 5 shows plot of average magnitude spectrum of noise and QBNE derived spectrum of noisy speech at two different quantiles. For effective noise cancellation, quantile spectrum that is closest to the average spectrum should be used. It was found that there is a close match between average spectrum of noise and QBNE derived spectrum of noisy speech at 2-5 percentile for frequencies below 900 Hz and at 20-30 percentile for frequencies above 900 Hz. Hence, a composite QBNE derived magnitude spectrum was used. The lower part, below 900 Hz, of the noise spectrum was obtained from 2-5 percentile spectrum and the upper part, above 900 Hz, was obtained from 20-30 percentile spectrum. For enhancement of alaryngeal speech, processing parameters as established earlier [8] were used: N=twice pitch period, $\alpha = 2$, $\beta = 0.001$, $\gamma = 1$.



(a) Recorded alaryngeal speech using NP-1 electrolarynx



(b) Enhanced speech, with averaged noise spectrum

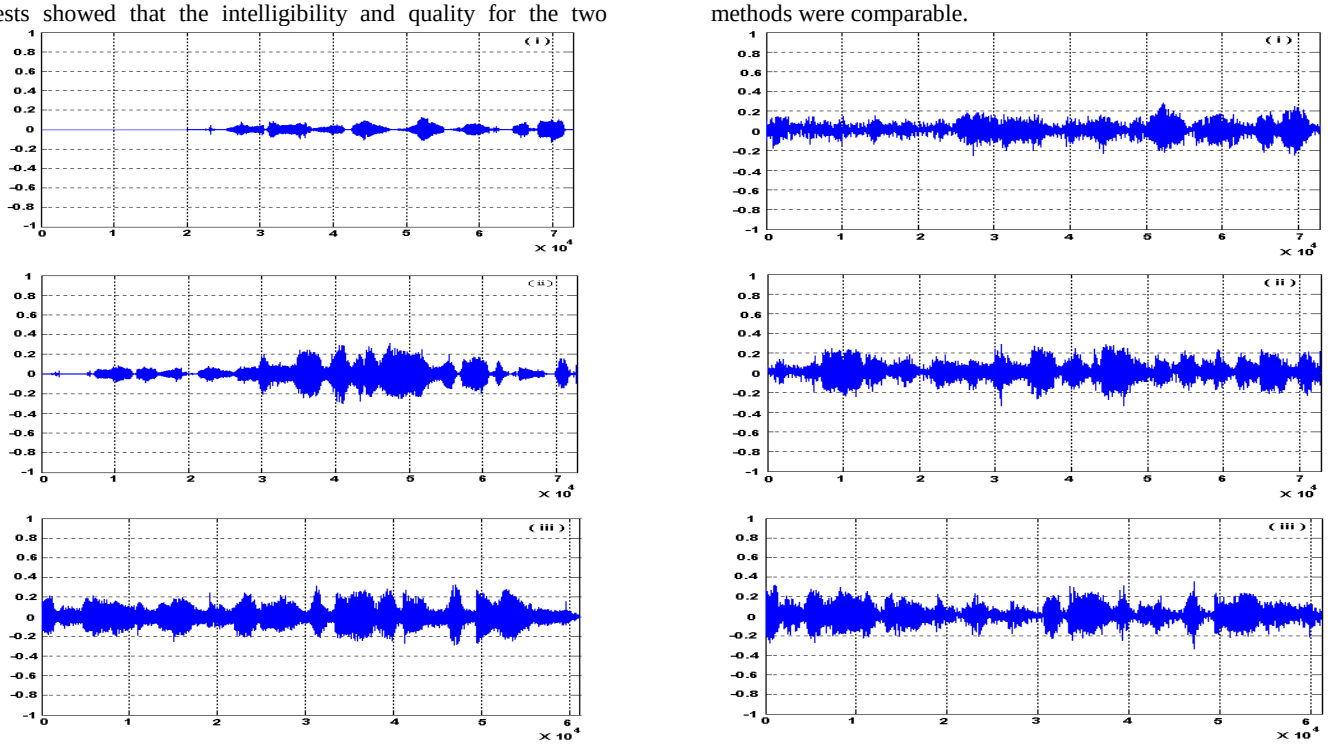


(c) Enhanced speech, with composite QBNE derived spectrum

Fig. 6. Recorded and enhanced speech, speaker: SP, material: silence followed by “Where are you going?”.

The enhancement method was tested over recordings of alaryngeal speech of short duration (~ a few second) and longer duration (~ 1 minute). In the beginning part of the recording, the speaker kept the lips closed. Background noise in the “silence” was used for obtaining average magnitude spectrum. Fig. 6 shows the results for shorter duration speech (4.8 sec). We see that that the results are comparable for both methods. Listening

tests showed that the intelligibility and quality for the two methods were comparable.



(a) with averaged noise spectrum (b) with composite QBNE derived spectrum

Fig. 7. Enhanced speech with two estimates of noise spectrum, speaker: SP, material: 3 question-answer pairs in English.

Fig.7 shows an example of enhancement with the two techniques for long duration recordings. It is seen that spectral subtraction using averaged noise spectrum from the "silence" part is not effective. There is over-subtraction at times, while under subtraction at other times. This can be attributed to variation in the characteristics of background noise. Spectral subtraction with QBNE provides a consistent noise subtraction.

The enhanced speech using QBNE based technique was of a better quality and intelligibility. This indicates that it was effective in obtaining a continuously updated estimate of noise spectrum for effective subtraction of background noise.

6. CONCLUSION

We have earlier reported [8] a single input technique for reduction of background noise in alaryngeal speech produced with a transcervical electrolarynx. It is based on application of spectral subtraction in a pitch-synchronous manner. The noise spectrum is estimated by averaging the magnitude spectra over several segments of background noise acquired with speaker keeping the lips closed. In this paper, we have presented an investigation for using quantile based noise spectrum estimation for continuous updating of noise spectrum. It is seen that for effective noise cancellation, a composite QBNE derived magnitude spectrum can be constructed from two quantile spectra: the lower part (below 900 Hz) of the noise spectrum from 2-5 percentile spectrum and the upper part (above 900 Hz) from 20-30 percentile spectrum. This method provides noise spectrum updates throughout speech/non-speech segments and results in effective reduction of background noise over extended periods.

For obtaining the QBNE derived composite noise spectrum, the optimum quantile values need to be investigated for different models of electrolarynx and for individual users. Use of frequency dependent quantile values, rather than just two quantile values in two frequency bands, may give further improvements without increasing computational load. Further the effect of the number of frames used in noise updating has to be investigated. The signal processing scheme for noise enhancement needs to be implemented as a real time DSP system and integrated with the electrolarynx for enhanced speech output.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, **Digital Processing of Speech Signals**, Englewood Cliffs, New Jersey: Prentice Hall, 1978.

[2] Y. Lebrun, "History and development of laryngeal prosthetic devices", **The Artificial Larynx**, Amsterdam: Swets and Zeitlinger, 1973, pp. 19-76.

[3] L. P. Goldstein, "History and development of laryngeal prosthetic devices", **Electrostatic Analysis and Enhancement of Alaryngeal Speech**, Springfield, Ill.: Charles C. Thomas Pub Ltd., 1982, pp. 137-165.

[4] Qi Yingyong and B. Weinberg, "Low-frequency energy deficit in electro laryngeal speech", **J. Speech and Hearing Research**, Vol. 34, 1991, pp. 1250- 1256.

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

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

[7] 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, 1979, pp. 1298-1308.

[8] P. C. Pandey, S. M. Bhandarkar, G. K. Bachher, P. K. Lehana, "Enhancement of alaryngeal speech using spectral subtraction", **Proc. 14th Int. Conf. Digital Signal Processing** (DSP 2002), Santorini, Greece, 2002, pp. 591-594.

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

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

[11] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering", **Proc. ICASSP, Vol**. 3, 2000, pp. 1875-1878.

[12] N. W. D. Evans and J. S. Mason, "Noise estimation without explicit speech, non-speech detection: a comparison of mean, median and model based approaches", **Proc. Eurospeech,** Vol. 2, 2001, pp. 893-896.

[13] 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", **Proc. 14th Int. Conf. Digital Signal Processing** (DSP 2002), Santorini, Greece, 2002, pp.985-988.

[14] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, **Numerical Recipes in C**, Cambridge, UK: Cambridge University Press, 1992.