QUANTILE BASED NOISE ESTIMATION FOR SPECTRAL SUBTRACTION OF SELF LEAKAGE NOISE IN ELECTROLARYNGEAL SPEECH

Santosh S. Pratapwar
EE Dept, IIT Bombay
Powai, Mumbai 400 076, India
(+91-22) 2576 4417
santosh@ee.iitb.ac.in

Prem C. Pandey
EE Dept, IIT Bombay
Powai, Mumbai 400 076
(+91-22) 2576 7445
pcpandey@ee.iitb.ac.in

ABSTRACT

Transcervical electrolarynx is of great help in verbal communication to a large number of laryngectomy patients. The quality of the electrolaryngeal speech is generally low because of low frequency spectral deficiency due to poor coupling, lack of short-time pitch control, and low voiced/unvoiced contrast. Its intelligibility suffers from the presence of self or background noise, caused by leakage of the acoustic energy from the vibrator. It has been shown earlier that the spectral subtraction technique, developed earlier for enhancement of noisy speech, can be applied in a pitch-synchronous manner for reducing the leakage noise in electrolaryngeal speech. This paper extends the method for more effective enhancement by using a quantile based continuously updated estimate of the noise spectrum from noisy speech.

Keywords: Artificial Larynx, Electrolarynx, Electrolaryngeal Speech Enhancement, Spectral Subtraction, Quantile Based Noise Estimation

1. INTRODUCTION

In normal speech production, the lungs provide the air stream, the vocal chords in the larynx provide the vibration source for the sound, and the vocal tract provides the spectral shaping of the resulting speech [14]. In some cases of disease and injury, the larynx is surgically removed by an operation known as laryngectomy, and the patient (often known as a laryngectomee) needs external aids to communicate.

An artificial larynx [10],[8] is a device used to provide excitation to the vocal tract, as a substitute to that provided by a natural larynx. The external electronic larynx or the transcervical electrolarynx is the widely used type of device. It is hand held and pressed against the neck. It consists of an electronic vibration generator, the vibrations coupled to the neck move up the vocal tract. Spectral shaping of the waveform by the vocal tract results in 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 lower coupling efficiency through the throat tissue [17]. The unvoiced segments generally get substituted by the voiced segments. In addition to these, the major problem is that the speech output has a background noise, which degrades the quality of the output speech considerably [10],[4].

Paper Identification Number: CI-2.2

This paper has been published by the Pentagram Research Centre (P) Limited. Responsibility of contents of this paper rests upon the authors and not upon Pentagram Research Centre (P) Limited. Individual copies could be had by writing to the company for a cost.

2. ELECTROLARYNGEAL SPEECH

A transcervical electrolarynx generally uses an electromagnetic transducer [1]. 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. Hence the background noise can be called as leakage or self-leakage noise.

Weiss *et al.* [16] and Barney *et al* [1] have reported detailed studies of perceptual and acoustical characteristics of electrolaryngeal speech. Speech-to-noise ratio (SNR), defined as the ratio of the average level of the vocal peaks in the electrolaryngeal speech (inclusive of background interference) and the level of radiated sound measured with the speaker's mouth closed, varied over 4-25 dB across speakers and devices. The leakage from the vibrator-tissue interface varied across speakers, and it significantly contributed to the background interference. The frequency and magnitude of spectral peaks in the leakage noise were speaker dependent. There was a 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.

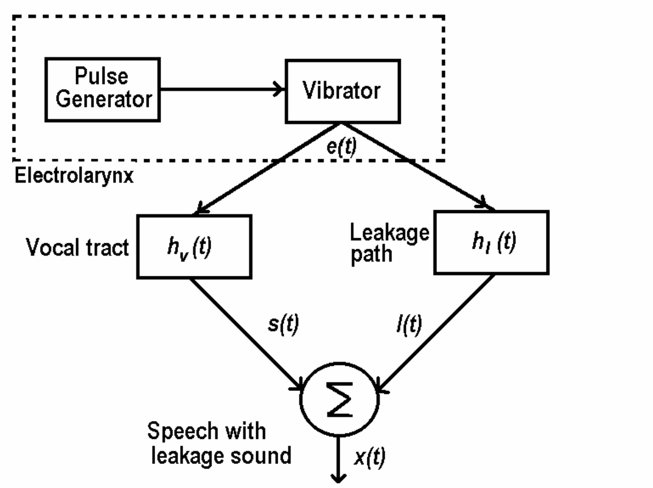


Figure 1: Model of background noise generation in transcervical electrolarynx [11].

A model of the leakage sound generation during the use of transcervical larynx [11] is shown in Fig.1. 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.1, are highly correlated.

Epsy-Wilson *et al.* [4] reported a technique for enhancement of electrolaryngeal 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 [4].

We have earlier reported a single-input noise cancellation technique based on spectral subtraction applied in a pitch synchronous manner [11]. In this technique, the noise spectrum is estimated by averaging the noise spectra over several segments of the self-leakage noise acquired with speaker keeping the lips closed. Because of variations in the noise characteristics, effective cancellation requires frequent acquisitions of noise. We have applied the use of quantile based noise spectrum estimation [6] for continuous updating of noise spectrum [12]. In this paper, we are presenting results of investigations with various types of QBNE based noise spectra for spectral subtraction.

After a review of the spectral subtraction technique for reducing the leakage noise [11], use of quantile based noise estimation is presented. This is followed by test results.

3. SPECTRAL SUBTRACTION FOR REDUCING LEAKAGE NOISE

Spectral subtraction technique is one of the important techniques for enhancing noisy speech [2],[3]. 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 electrolaryngeal speech, speech signal and leakage interference are not uncorrelated. With reference to Fig.1, 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

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

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 spectrum 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

$$|Y_n(k)|^2 = |X_n(k)|^2 - |L(k)|^2 \tag{6}$$

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

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

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 [2] to reduce spectral excursions is used.

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

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* [2] to the spectral subtraction algorithm is the addition of exponent factor γ in place of 2 for subtraction.

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

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 [2]. A schematic of the modified spectral subtraction algorithm is shown in Fig.2. Optimal values for reduction of background noise as reported by Pandey *et al* [11] are: window length= twice the pitch period, spectral subtraction factor $\alpha = 2$, spectral floor factor $\beta = 0.001$, and exponent factor $\gamma = 1$.

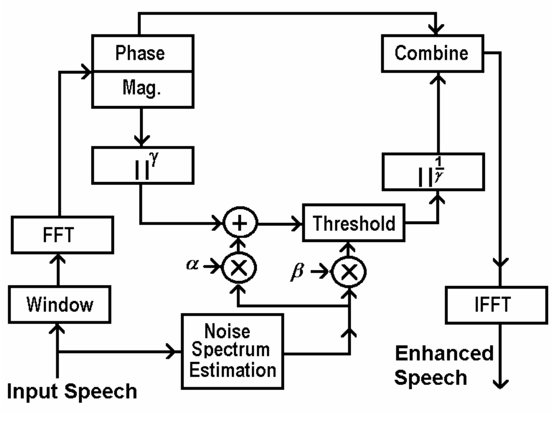


Figure 2: Spectral subtraction scheme [11].

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 background 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 updating of the estimated noise spectrum is required. Recursive averaging of spectra during silence segments may be used for noise spectrum estimation [3],[2]. However, speech/silence detection in electrolaryngeal speech is rather difficult. Quantile-based noise estimation (QBNE) technique [6] does not need speech/non-speech classification and we have investigated its use for noise estimation in electrolaryngeal speech.

4. INVESTIGAION WITH QBNE

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 [15],[5],[6]. 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 the 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 [13] was implemented. It gives spectral values for a quantile for each frequency sample; and these values can be used for quantile based noise estimation continuously [12].

The recordings were done with the microphone positioned at the center between the mouth and the artificial larynx position. During first 2 s, speaker kept the lips closed, and the recorded speech contained only noise. For training the QBNE method, a segment of speech of approximately 2 s is taken and quantile values are found for all frequency samples, such that they give average power spectrum of noise. Those quantile values were used as matched quantile values to be used for estimation of noise for subsequent speech signal. As the quantile values were found to have large variation over frequency, use of low pass filtered estimate of the quantile values was also investigated. The results were compared with those obtained by using median (50 percentile) estimated noise spectrum.

5. TEST RESULTS

Quantile based noise estimation technique was used for speech enhancement of electrolaryngeal 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.

Figure 3 shows waveforms and spectrogram of the self leakage noise produced by the electrolarynx, and three vowels /*a*/, /*i*/, and /*u*/ produced using the same device. It can be seen that the leakage noise is comparable in magnitude to output speech, and it has a vowel-like formant structure. Further, the spectrogram shows that the spectrum is short-term stationary. This characteristic is very different from that of other types of noises that degrade speech signal. Fig.4. shows the spectrum of the noise and that of vowel /*i*/ spoken using the same device as excitation source.

Analysis was carried out to obtain quantile values as a function of frequency, which will match the quantile based spectrum of noisy speech to the averaged spectrum of noise. These quantile values were used for estimation of noise spectrum during use mode. Quantile spectra of noisy speech were obtained using 55 frames with 50% overlap, i.e. segment length of 0.6 s contributed to updating. The matched quantile values showed a variation over 45-85, and changed with speech segments. Hence it was decided to use a smoothened version of quantile values. This was carried out by using a 9 frequency sample averaging (averaged over 810 Hz). Further, the matched quantile values were averaged over all the frequency samples and the averaged quantile was also used for noise estimation during use mode. Fig.5. shows plot of matched quantile values obtained from training mode, smoothened matched quantile values, and averaged quantile value.

Each of the above three quantile plots were used for estimation of noise spectrum, and these are shown in Fig.6. It is seen that noise spectrum estimated using averaged quantile value does not match with the averaged noise spectrum. The noise spectrum estimated from different speech segments using the smoothened quantile plot generally matched with the averaged spectrum.

Fig.7 shows a recording of a question-answer pair and enhancements carried using different estimates of noise. All the enhancement results shown here were obtained using $\alpha = 2$, $\beta = 0.01$, and $\gamma = 1$. In the unprocessed speech (Fig. 7a), the initial 1.7 s segment was the leakage noise (sound produced with the speaker's mouth closed). The subsequent 2.1 s segment was the noisy speech. Enhancement using spectral subtraction (Fig. 7b-f) showed a significant reduction in noise. It permitted scaling of the signal by a factor of 4, without causing saturation. Enhanced speech with averaged estimate of noise (Fig. 7b) totally eliminated the leakage noise, but it resulted in significant over subtraction during noisy speech. The other four enhancements (Figs. 7c-f) used noise spectrum obtained with QBNE and different quantile estimates. These did not totally eliminate the noise during non-speech segment. Median estimated spectrum resulted in over subtraction, actually more than for average estimated. Averaged match quantile estimated spectrum showed less over subtraction and speech clipping. Spectra estimated by matched quantile and smoothened matched quantile showed almost similar level of noise reduction and signal subtraction. However, an informal listening test indicated that the smoothened matched quantile resulted in better quality speech output.

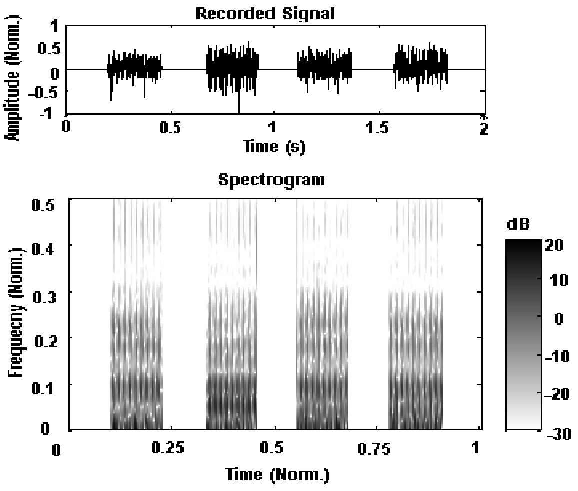


Figure 3: Waveforms and spectrogram of self leakage noise of electrolarynx and three vowels /*a*/, /*i*/, and /u/ produced using the same electrolarynx.

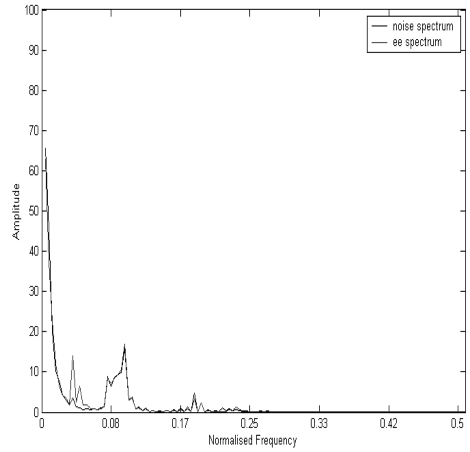


Figure 4: Average power spectrums of self leakage noise, and that of vowel /*i*/ produced using the same device.

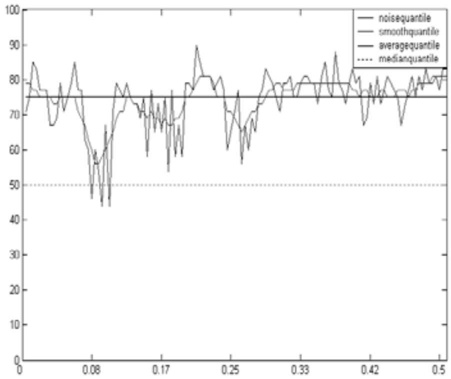


Figure 5: Plot of matched quantile values, smoothed quantile values, and average of quantile values.

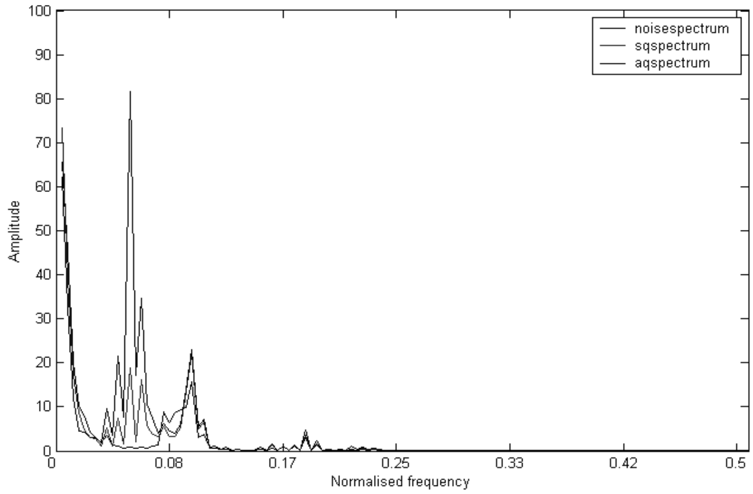
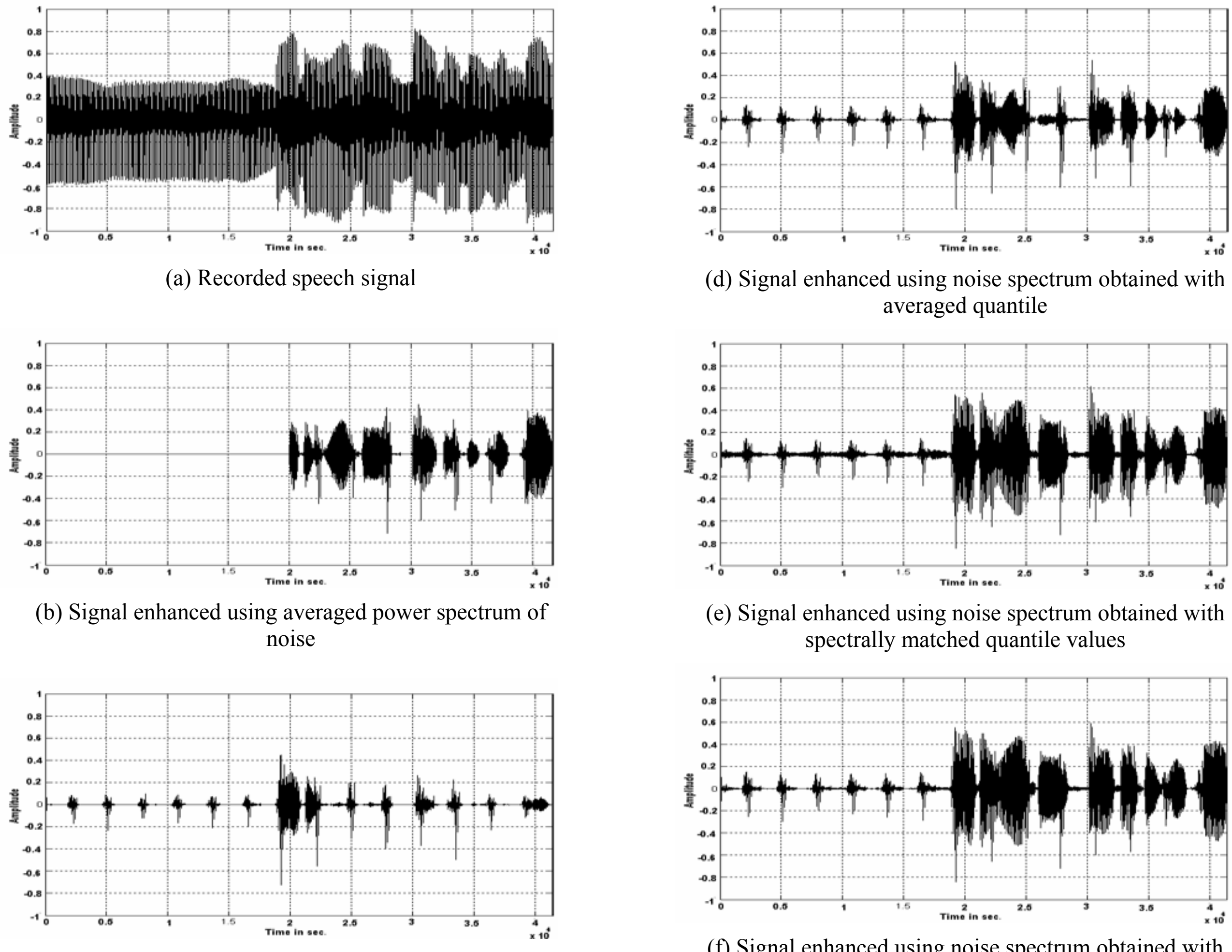


Figure 6: Average power spectra obtained using optimal quantile values and average quantile values



(a) Recorded speech signal

(d) Signal enhanced using noise spectrum obtained with averaged quantile

(b) Signal enhanced using averaged power spectrum of noise

(e) Signal enhanced using noise spectrum obtained with spectrally matched quantile values

(c) Signal enhanced using noise spectrum obtained with median.

(f) Signal enhanced using noise spectrum obtained with smoothened quantile values.

Figure 7: Recorded and enhanced speech with five different estimates of noise spectrum. Speaker: SP, material: question-answer pair in English, "*What is your name? My name is Santosh*".

6. CONCLUSIONS

We have earlier reported application of spectral subtraction in a pitch synchronous manner for enhancement of electrolaryngeal speech, by using an averaged estimate of leakage noise spectrum [11]. Quantile based noise estimation was later applied for obtaining a continuously updated estimate of noise spectrum without speech vs non-speech detection [12]. In this paper, we have presented an investigation involving different quantile estimates for finding an estimate of noise spectrum for providing significant reduction in leakage noise without appreciable reduction or clipping of signal. It is found that noise spectrum estimated with matched quantile values smoothened along frequency samples meets this requirement. A real time implementation of the technique may be incorporated as part of the artificial larynx for a better quality speech.

7. REFERENCES

[1] Barney, H. L., Haworth, F. E., and Dunn, H. K. An experimental transistorized artificial larynx Bell Systems Tech. J., 38, 6 (1959), 1337-1356.

[2] Berouti, M., Schwartz, R., and Makhoul, J. Enhancement of speech corrupted by acoustic noise. In Proc. ICASSP '79, (1979), 208-211.

[3] Boll, S. F. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. ASSP '79**,** 27, 2 (1979), 113-120.

[4] Espy-Wilson, C. Y., Chari, V. R., and Huang, C. B., Enhancement of electrolaryngeal speech by adaptive filtering. In Proc. ICSLP '96, (1996), 764-771.

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

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

[7] Evans, N. W. D., and Mason, J. S. Time-frequency quantile-based noise estimation. In Proc. EUSIPCO '02 (2002).

[8] Goldstein, L. P. History and development of laryngeal prosthetic devices. Electrostatic Analysis and Enhancement of Electrolaryngeal Speech. Charles C. Thomas Pub Ltd, Springfield, Mass., 1982, 137-165.

[9] Houwu, B., and Wan, E. A. Two-pass quantile-based noise spectrum estimation. In Oregon Health and Science University (OHSU) Technical Report, (2002). http://citeseer.nj.nec.com/560702.html

[10] Lebrun, Y. History and development of laryngeal prosthetic devices, In The Artificial Larynx**.** Swets and Zeitlinger, Amsterdam, (1973), 19-76.

[11] Pandey, P. C., Bhandarkar, S. M., Bachher, G. K., and Lehana, P. K. Enhancement of electrolaryngeal speech using spectral subtraction. In Proc. Int. Conf. Digital Signal Processing DSP 2002 (Santorini, Greece, 2002), 591-594.

[12] Pratapwar, S. S., Pandey, P. C., and Lehana, P. K. Reduction of background noise in alaryngeal speech using spectral subtraction with quantile based noise estimation. In Proc. of 7th World Multiconference on Systemics, Cybernetics and Informatics SCI, (Orlando, USA, 2003).

[13] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. Numerical Recipes in C. Cambridge University Press, Cambridge, (1992).

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

[15] Stahl, V., Fischer, A., and Bippus, R. Quantile based noise estimation for spectral subtraction and Wiener filtering. In Proc. ICASSP '00, (2000), 1875-1878.

[16] Weiss, M., Komshian, G. Y., and Heinz, J. Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. J. Acoust. Soc. Am., 65, 5 (1979), 1298-1308.

[17] Yingyong, Qi., and Weinberg, B. Low-frequency energy deficit in electro laryngeal speech, J. Speech and Hearing Research, 34 (1991), 1250-1256.