Proc. 14th Int. Conf. Digital Signal Processing (DSP 2002), Santorini, Greece, 2002, pp. 591-594

ENHANCEMENT OF ALARYNGEAL SPEECH USING SPECTRAL SUBTRACTION

Prem C Pandey[1], Santosh M. Bhandarkar[1], Gurmeet K Bachher[2], Parveen K Lehana[1]

[1]EE Dept, IIT Bombay, Powai Mumbai, India
pcpandey@ee.iitb.ac.in
[2]Speech Therapy Dept, Tata Memorial Hospital, Parel Mumbai, India

Abstract: The transcervical electrolarynx is of great help to persons who cannot use their natural voice production mechanism. The device is held against the neck, and the vibrations generated move up the vocal tract to produce useful speech. The presence of background noise, caused by leakage of the acoustic energy from the vibrator, degrades the resulting speech. After formulating a theoretical basis, modified spectral subtraction method is used for speech enhancement. Average magnitude spectrum of noise, obtained with lips closed in training mode, is subtracted from the magnitude spectrum of the noisy speech and the signal is reconstructed using the original phase spectrum. It is observed that effective noise cancellation is obtained, if the noise estimation and subtraction is done using 2-pitch frames. The set of optimal values for processing parameters of the method are empirically obtained.

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 [1]. 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 [2],[3] 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 [4]. 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 [2],[5].

After formulating a theoretical basis, we present a technique of enhancing alaryngeal speech (produced using electrolarynx) by using spectral subtraction method for reducing the background noise.

2. BACKGROUND NOISE IN ALARANGEAL SPEECH

A transcervical electrolarynx generally uses an electromagnetic transducer [6]. The steady background noise is generated due to the leakage of the vibrations produced by the vibrator membrane/plate. Its front end is coupled to the neck, while the back end is coupled to the air in the instrument housing. Leakage of the acoustical energy from the housing to the air outside is responsible for the production of the noise. This noise is present even if the speaker's lips are closed. Leakage of vibrations from the front end of the vibrator membrane/plate due to improper coupling of the vibrator to the neck tissue also contributes to the background noise. Fig. 1 shows a model of the leakage sound generation during the use of the transcervical electrolarynx. The vibrations generated by the vibrator diaphragm have two paths. The first path is through the neck on to the vocal tract and the second one is through the surroundings known as the leakage path. The impulse response of the first path $h_v(t)$ is dependent on various factors such as the length and configuration of the vocal tract of the speaker, the place of coupling of the vibrator, the amount of coupling, etc. The excitation pulse $e(t)$ when passed through the vocal tract filter delivers the useful speech $s(t)$. The leakage of the vibrations is perceived as the component $l(t)$, which gets added to the useful speech $s(t)$, thereby deteriorating its intelligibility.

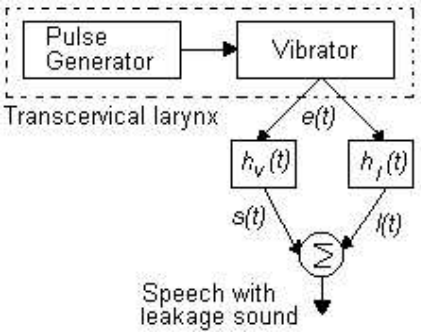


Fig. 1. Model of background noise generation due to leakage of vibrations in a transcervical electrolarynx.

Effective acoustic shielding of the vibrator should reduce the noise. However, studies have shown that shielding 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 source. However such vibrators have poor efficiency. Further vibrator design cannot help in reducing the leakage from the vibrator to tissue interface.

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 SNR' s. 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 SNR' s, 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 SNR' s. However, the background noise reduced the identification of consonants.

Noise reduction can be done by estimating the noise present in the signal and subtracting it from the noisy signal. The main problem in attempting signal processing techniques is that the noise is highly correlated to speech signal as they result from the same excitation source as shown in Fig. 1. A technique for enhancement of alaryngeal speech was reported by Espy-Wilson et al. [5], using two-input adaptive filtering based on LMS algorithm. For the sound produced by an artificial larynx, the noisy signal and the noise reference are strongly correlated for voiced sounds, because the vibrations are derived solely from the artificial larynx. However, in case of consonants (unvoiced segments, to be more precise) the correlation gets weaker on account of the vocal excitation being caused by the turbulence at constrictions. During the vocal sounds, if the adaptation is allowed then the cancellation will result in an output that contains no information at all. Authors reported that during the non-sonorant or the low energy segments, the noise cancellation was effective, and most of the background noise was cancelled. However during the sonorant intervals there was an improvement in the output quality, though the background noise was not removed fully. The intelligibility of the processed output was better than that of its unprocessed counterpart.

As shown in Fig. 2, a system for enhancement of alaryngeal speech based on a single input signal processing technique is envisaged. It has two modes of operation, namely the training mode and the use mode. In the training mode, the speaker keeps his lips closed and noise forms the input, while during the use mode, the noisy speech forms the input. The algorithm is based on estimation of magnitude spectrum of background noise and spectral subtraction during the use mode, with the objective of reducing background noise and enhancing alaryngeal speech.

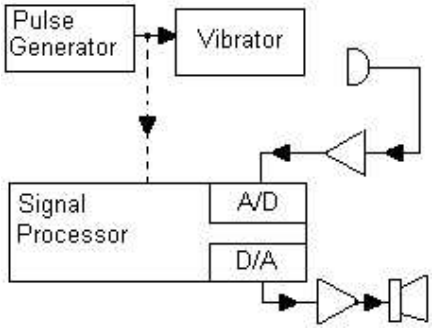


Fig. 2. Real-time implementation scheme of noise cancellation with spectral subtraction algorithm.

3. SPECTRAL SUBTRACTION TECHNIQUE

Spectral subtraction method is one of the important techniques for enhancing noisy speech [8],[9]. The basic assumption made is that the clean speech and the noise are uncorrelated, and therefore the magnitude spectrum of the noisy speech signal equals the sum of magnitude spectrum of noise and clean speech. In case of alaryngeal speech, speech signal and background interference are not uncorrelated. With reference to Fig. 1, let $x(n)$ be the noisy speech, $h_v(n)$ be the impulse response of the vocal tract, $h_l(n)$ be the impulse response of the leakage path, and $e(n)$ be the excitation signal. The noisy speech signal is given as

$$x(n) = s(n) + l(n) \quad (1)$$

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

$$s(n) = e(n) * h_v(n) \quad (2)$$

$$l(n) = e(n) * h_l(n) \quad (3)$$

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

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

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

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

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

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

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

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 [9] 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. [9] 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. The normalization factor G is given as

$$G = \{(|X_n(k)|^2 - |L(k)|^2)/\, |Y'_n(k)|^2\}/\gamma \quad (10)$$

if the current value of G is less than its previous value. If the current value is greater than the previous value, the previous value of G is retained.

4. TEST RESULTS

After implementing the modified spectral subtraction method, as shown in Fig. 3, investigations were carried out for establishing the window size and optimal values of α, β, and γ for enhancement of alaryngeal speech.

The recordings were done with the microphone positioned at the center between the mouth and the artificial larynx position. The total duration of the recording was 5 s, at a sampling rate of 11.025 kHz, during first 2 s, speaker kept the lips closed, and the recorded speech contained only noise. Investigations were done using electrolarynx model NP-1 (manufactured by NP Voice, India).

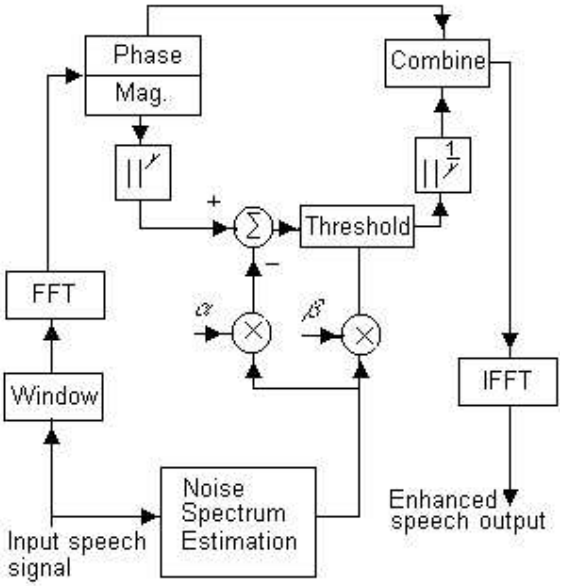
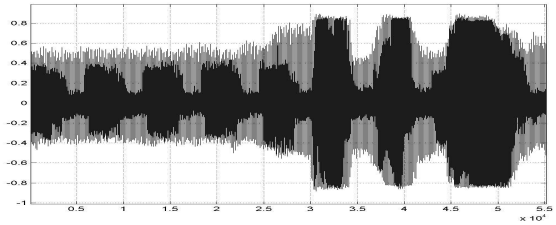


Fig. 3. Modified spectral subtraction algorithm [9].

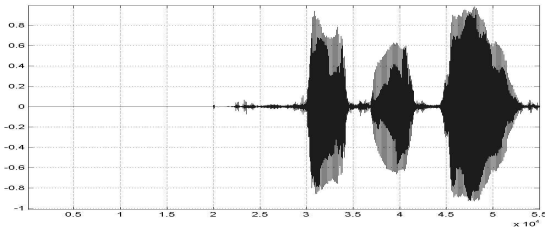
The vibrator of the electrolarynx had a fixed pitch of 90.3 Hz, i.e. pitch period of 122 samples. For studying the effect of the window size on the output speech, various window sizes, ranging from 60 to 1500 samples, were experimented with. For window size less than the pitch period, the output was highly degraded. As the window size was increased beyond the pitch period, the output quality got better. However, with the window size much larger than twice the pitch period, there was slight degradation in the output on account of warbling. The best results were obtained at window size of 244, corresponding to two pitch periods. Therefore this

window length was used for seeing the effect of other parameters.

The effects of different values for subtraction factor α on the output were investigated. For $\alpha < 1$, there was considerable amount of noise present in the output. For $\alpha > 1$ the results contained a small amount of broadband noise. With increasing α, the broadband noise in the output decreased considerably. For very large α, some portion of the useful speech information is lost. The range of α from 2 to 5 was considered appropriate. The effect of the spectral floor factor β values on the output was tested. The value of β was varied between 0 and 1, with α set at 2. Increasing the value of β reduces the residual noise. Increasing the β value beyond 0.1 introduces broadband noise in the output. The best results were obtained with $\beta = 0.001$. Finally the effects of different values of exponent factor γ on the output were investigated, waveforms for phrase "plug and play" are shown in Fig. 4. For $\gamma < 1$, good results were obtained. However there was a need for normalization since the amplitude of the output was low. With $\gamma = 1$, the results obtained were better. Increasing γ beyond 1 gave inferior output quality. The value of $\gamma = 1$ seemed to be an appropriate value. In summary, these tests indicated that for a good balance between perceptual quality and speech-to-noise ratio, the parameters to be used are: $\alpha = 2$, $\beta = 0.001$, and $\gamma = 1$.



Unprocessed speech generated using NP-1 electrolarynx



Processed waveform, spectral subtraction method, N = 244, $\alpha = 2$, $\beta = 0$, $\gamma = 1$

Fig. 4. Unprocessed and processed speech "plug and play" generated using NP-1 electrolarynx.

5. CONCLUSION

A single input processing technique using modified spectrum subtraction has been developed and tested for enhancement of alaryngeal speech. It has been found that the best results are obtained when the processing window length equals two pitch periods. Normally in an artificial larynx, the pitch period is fixed, hence the signal processor can use a fixed pitch value. In case of variable pitch control, the signal processing block can be given trigger pulses from the pulse generator, and detection of glottal pulses by signal processing may not be required.

There are three processing parameters in this technique: subtraction factor α, spectral floor factor β, and exponent factor γ. On the basis of theoretical formulation, we should select $\alpha = 1$ (complete subtraction of magnitude speech of noise), $\beta = 0$ (zero noise floor), and $\gamma = 2$ (subtraction of power spectrum). Experimental investigation indicated that for effective noise suppression, we have to use over-subtraction with $2 < \alpha < 5$. For a reasonable balance between musical and broadband noises, we should use $0.001 < \beta < 0.01$. Best perceptual quality was obtained with $\gamma = 1$. In general, different settings of the parameters may be needed for different devices and users.

The effect of positioning the window, in synchronization with the impulse, needs to be investigated. In the method investigated, the phase spectrum of the noisy speech was retained and coupled with the "cleaned" magnitude spectrum for obtaining processed speech for each window segment. It is expected that quality can be better if phase spectrum also is noise-free. Towards this, it can be noted that vocal tract as well as leakage path can be modeled as minimum phase, because of their passive nature. The resynthesis of the phase from the "cleaned" magnitude spectrum for further enhancing the speech synthesis needs to be investigated.

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, pp. 19-76, 1973.

[3] L. P. Goldstein, "History and development of laryngeal prosthetic devices," *Electrostatic Analysis and Enhancement of Alaryngeal Speech*, pp. 137-165, year not known.

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

[5] C.Y. Espy-Wilson, V.R. Chari, and C.B. Huang, "Enhancement of alaryngeal speech by adaptive filtering," *Proc. ICSLP 96*, pp. 764-771, 1996.

[6] 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, Nov. 1959.

[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, pp. 1298-1308, May 1979.

[8] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE*, ASSP-27, pp. 113-120, 1979.

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