USE OF HARMONIC PLUS NOISE MODEL FOR REDUCTION OF SELF LEAKAGE IN ELECTROALARYNGEAL SPEECH

Parveen K. Lehana
Dept of Electronics
University of Jammu,
Jammu 180 006, India
(+91-191) 2453079
lehana@iitb.ac.in

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

Rockey Gupta
Dept of Electronics
University of Jammu,
Jammu 180 006,India
(+91-191) 2453079
rockegupta@yahoo.com

ABSTRACT

Artificial larynx is an assistive device for providing excitation to vocal tract as a substitute to a dysfunctional or removed larynx. The speech generated by electrolarynx, an external vibrator held against the neck tissue, is not natural and most of the time is unintelligible because of the improper shape of the excitation pulses and presence of a background noise caused by sound leakage from the vibrator. The objective of this paper is to enhance the intelligibility of electrolaryngeal speech by reducing the background noise using harmonic plus noise model (HNM). The alaryngeal speech and the leakage signal are analyzed using HNM and average harmonic spectrum of the leakage noise is subtracted from the harmonic magnitude spectrum of the noisy speech in each frame. HNM synthesis is carried out retaining the original phase spectra. Investigations show that the output is more natural and intelligible as compared to input speech signal and the enhanced signal obtained from spectral subtraction without HNM analysis and synthesis.

Categories and Subject Descriptors

[**Informatics**]: *Speech / audio signal processing, biomedical engineering applications.*

General Terms

Algorithms, Experimentation.

Keywords

Enhancement, alaryngeal speech, harmonic plus noise model.

1. INTRODUCTION

A very flexible acoustic model for speech synthesis is based on sinusoidal modeling of speech – the waveform is represented as summation of a finite number of sinusoids with arbitrary amplitudes, frequencies, and phases [8]. Harmonic plus noise model, a variant of sinusoidal model, requires smaller number of parameters and is computationally efficient. Time and frequency scaling is also relatively easy [14]. As speech output from the vocal tract is given by the convolution of the excitation and the impulse response of the vocal tract, we can use the following equations to obtain the output of the vocal tract.

Let the excitation to the vocal tract is given by

$$e(t) = \sum_{k=1}^{K(t)} \exp(j\phi_k(t)) \qquad (1)$$

and vocal tract transfer function is written as

$$H(f;t) = G(f;t)\exp[j\varphi(f;t)] \qquad (2)$$

Then the output speech can be given by:

$$s(t) = \sum_{k=1}^{K(t)} G(f_k(t);t)e^{j(\phi k(t)+\varphi(fk(t);t))} \qquad (3)$$

which can be simplified to

$$s(t) = \sum_{k=1}^{K(t)} A_k(t)e^{i\theta k(t)} \qquad (4)$$

where instantaneous phase of the k^{th} harmonic $\theta_k(t)$ equals the sum of the excitation phase and system phase

$$\theta_k(t) = \phi_k(t) + \varphi(f_k(t);t)) \qquad (5)$$

In equation (4) the amplitude of the k^{th} harmonic of the fundamental $A_k(t)$, is equal to the product of the amplitudes of the corresponding harmonics of the spectrum of the excitation and the transfer function of the vocal tract. HNM uses the properties of equations (4) and (5) for calculating parameters.

HNM divides the spectrum of the speech in two sub bands: one is modeled with harmonics of the fundamental and the other is simulated using random noise. The frequency, which partitions the speech spectrum in these two bands, is known as maximum voiced frequency (F_m) and varies from frame to frame. In natural speech, the noise part is contributed by non-periodic components as fricative or aspiration noise and period-to-period variation of the glottal excitation, etc. As most of the non-periodic components from the speech are removed in harmonic part, it may be expected that synthesis from harmonic part is less susceptible to noise mixed in the input speech. In HNM synthesis, the analysis/synthesis is performed at each glottal closure instant (GCI) [14], and hence it is a pitch synchronous technique. This is well suited for high quality speech output, but we need to estimate the GCIs precisely. We have explored the HNM for enhancement of electrolaryngeal speech. The generation of the alaryngeal speech is discussed in the following paragraph.

Paper Identification Number: CI-3.1

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.

In normal speech production system, the lungs provide the air stream; the vocal chords in the larynx provide the spectral shaping of the resulting speech [13]. In some cases where the larynx has been removed due to cancer or injury, the patient needs some aid to communicate. An artificial larynx [6],[10] is an assistive device, which is used to provide the excitation to the vocal tract, as a substitute to that provided by natural larynx. It is a vibrating source to provide air vibrations similar to the normal larynx.

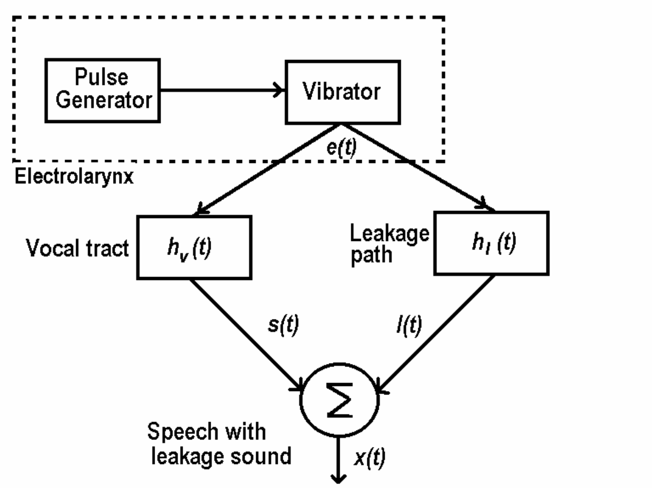


Figure 1. Model of production of electrolaryngeal speech [10].

The most common type in use at present is transcervical electrolarynx. It is a small hand-held vibrator, which is pressed against the neck so that vibrations are transmitted into the vocal tract and the speech can be produced. A transcervical electrolarynx generally uses an electromagnetic transducer [1]. The steady background noise is generated due to the leakage of the vibrations produced by the vibrator membrane/plate and the amount depends on length & configuration of the vocal tract, the place and amount of coupling, etc. Fig. 1 shows a model of the production of electrolaryngeal speech. The excitation pulse $e(t)$ convolved with the impulse response $h_v(t)$ of the vocal tract delivers the speech signal $s(t)$. The impulse response $h_v(t)$ depends 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 component $l(t)$ is leakage of the acoustical energy from the housing to the air outside, and is of background noise. This noise is present even if the speaker's lips are closed. Leakage of the vibrations from the front end of the vibrator membrane/plate is due to improper coupling of the vibrator to the neck. Leakage from the back end is caused by inadequate shielding of the sound generated within the vibrator. The leakage component *l(t)* gets added to useful speech and deteriorates its intelligibility.

Spectral subtraction technique [2],[3] is often used for enhancement of speech corrupted by uncorrelated noise. However, as we see in Fig. 1, the speech and leakage components originate from the same excitation and hence are related. It has been earlier shown [10] that a pitch synchronous application of spectral subtraction can be employed for enhancement of electrolaryngeal speech. As a further work in this direction, we have investigated the use of HNM based analysis and synthesis for the enhancement of alaryngeal speech. Investigations show that the output is comparatively natural, intelligible and has less background noise as compared to the input.

2. HNM ANALYSIS / SYNTHESIS

In HNM based synthesis, the speech signal is assumed to be composed of two parts: harmonic and noise

$$s(t) = s'(t) + n'(t) \qquad (6)$$

The harmonic part accounts for the quasi-periodic components of the speech signal while the noise part is responsible for non-periodic components: fricative or aspiration noise, period-to-period variation of the glottal excitation, etc. The frequency separating the two bands is called maximum voiced frequency F_m [9],[11]. In the lower band, the signal is represented only by harmonically related sine waves with slowly varying amplitudes and frequencies

$$s'(t) = \mathrm{Re} \sum_{l=0}^{L(t)} a_l(t) \exp\{j[\int_0^t l w_0(\sigma) d\sigma + \theta_l]\} \qquad (7)$$

where $a_l(t)$ and $\theta_l(t)$ are the amplitude and phase at time t of the l^{th} harmonic, ω_0 is fundamental frequency and $L(t)$ is the time-varying number of harmonics included in the harmonic part. In the upper band, the signal is represented as noise $n'(t)$, modeled as a white Guassian noise $b(t)$ filtered by a time-varying all-pole normalized filter $h(\tau\,;\,t)$ and amplitude modulated by an envelope function $w(t)$

$$n'(t) = w(t)[h(\tau;t) * b(t)] \qquad (8)$$

Analysis scheme is shown in Fig. 2. Speech signal is applied to the voicing detector, which declares the frames either voiced or unvoiced. As the analysis and synthesis in HNM is pitch-synchronous, it is necessary to estimate glottal closure instants (GCIs) precisely. For each voiced frame the maximum voiced frequency (F_m) is calculated. The analysis frame is taken twice the local pitch period. This voiced frame of the speech is analyzed at each GCI for calculating amplitudes and phases of all the pitch harmonics up to F_m. GCIs are obtained by processing the speech signal using Childers and Hu's algorithm [4].

Synthesis scheme is shown in Fig. 3. As the analysis parameters were obtained on frame-by-frame basis, they are interpolated before synthesis for obtaining the parameters values at each sample. Linear interpolation was used for this purpose. Phase unwrapping is carried out before interpolation of the phase values. Adding the synthesized harmonic and synthesized noise part gives the synthesized speech. It is to be noted that the harmonic synthesis is used in the analysis of the speech for obtaining the noise part as shown in Fig. 2.

For calculating the noise parameters, the synthesized voice part of speech is subtracted from the original speech. Noise part is analyzed for obtaining LPC coefficients and energy envelope. The length of the analysis window for noise part is taken as two local pitch periods for both voiced and unvoiced frames. For voiced frames the local pitch is the pitch of the frame itself. For unvoiced frames the pitch of the last voiced frame is taken.

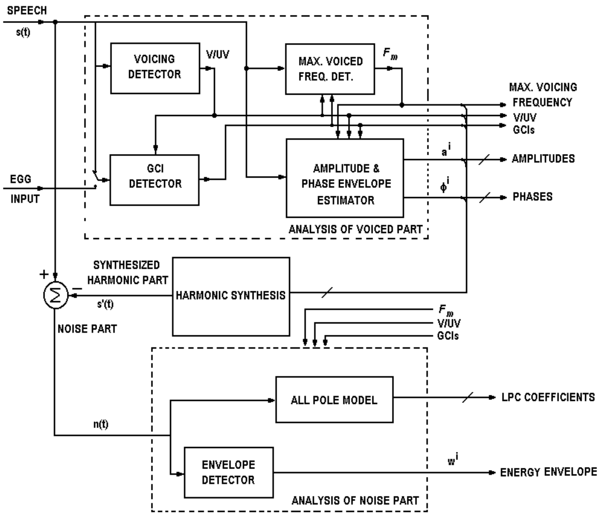


Figure 2. Analysis of speech using HNM [7].

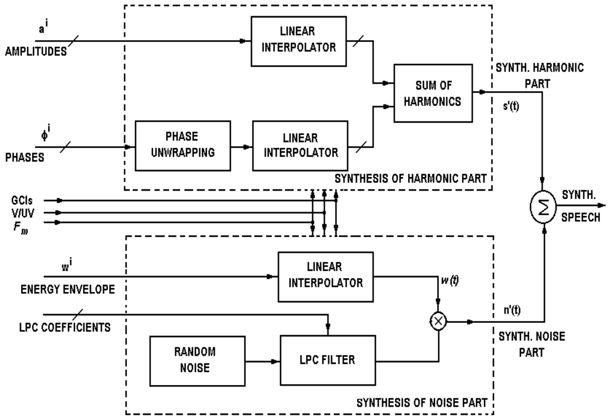


Figure 3. Synthesis of speech using HNM [7].

3. SPECTRAL SUBTRACTION WITH HNM

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

$$x(n) = e(n) * h_v(n) + e(n) * h_l(n) \qquad (9)$$

Taking short-time Fourier transform on either side of (9), we get

$$X_n(e^{jw}) = E_n(e^{jw})[H_{vn}(e^{jw}) + H_{\mathrm{ln}}(e^{jw})] \qquad (10)$$

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

$$|X_n(e^{jw})|^2 = |E_n(e^{jw})|^2 \, [|H_{vn}(e^{jw})|^2 + |H_{\mathrm{ln}}(e^{jw})|^2] \qquad (11)$$

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

$$|X_n(e^{jw})|^2 = |L_n(e^{jw})|^2 = |E_n(e^{jw})|^2 |H_{\mathrm{ln}}(e^{jw})|^2 \qquad (12)$$

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

$$|Y_n(k)|^\gamma = |X_n(k)|^\gamma - \alpha |L(k)|^\gamma$$

$$|Y'_n(k)|^\gamma = |Y_n(k)|^\gamma \quad \text{if } |Y_n(k)|^\gamma > \beta |L(k)|^\gamma$$

$$= \beta |L(k)|^\gamma \text{ Otherwise} \qquad (13)$$

where k is the frequency index and n is the frame index.

In case of electrolaryngeal speech, the excitation pitch remains basically constant. Hence the background noise and the speech signal both can be adequately modeled by the harmonic part alone of the HNM. Hence we can explore the use of HNM analysis / synthesis in spectral subtraction for reduction of background noise in electrolaryngeal speech.

4. METHODOLOGY

The spectral subtraction technique was implemented using Harmonic plus noise model (HNM). The scheme for analysis and synthesis is shown in Fig. 4. Non-speech segments are analyzed for obtaining average harmonic spectrum. During speech, the noisy speech average harmonic spectrum is subtracted from the harmonic spectrum of each frame. The phase spectrum of noisy speech is retained for resynthesis. For subtraction of the harmonic spectrum the modified spectral subtraction method [2] to reduce spectral excursions is used. For HNM based analysis / synthesis the index k in equation (13) represents the harmonic index. The values of the parameters used were: α = 1, β = 0.1, and γ = 1. The modified parameters are used for HNM synthesis for generating enhanced speech.

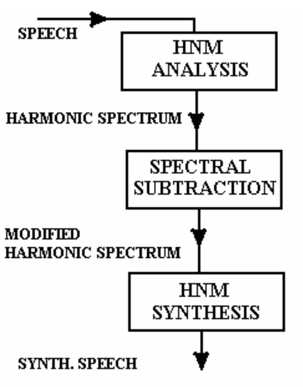


Figure 4: Analysis/synthesis scheme for HNM based spectral subtraction.

5. RESULTS

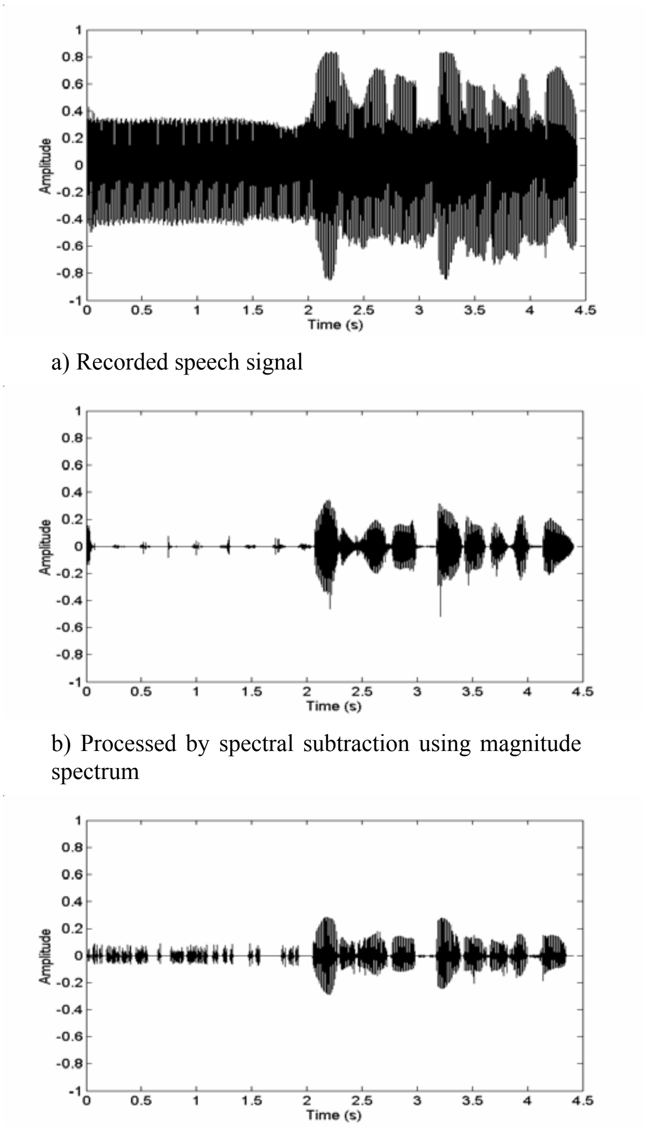
Electrolaryngeal speech was recorded using a B & K 4176 microphone with electrolarynx model NP-1 (manufactured by NP Voice, India). Fig. 5 shows the waveforms for the recorded and processed speech. First segment (2 seconds) corresponds to leakage noise and the subsequent is noisy speech. Averaged harmonic spectrum was computed over this period. This spectrum was used for spectral subtraction over the entire duration (including the initial non-speech part). For comparison, the waveform enhanced by spectral subtraction using DFT derived magnitude spectrum, and $\alpha = 2$, $\beta = 0.001$, and $\gamma = 1$. An example of enhancement by the two spectral subtraction methods, (i) DFT derived spectrum (ii) HNM based harmonic spectrum, is shown in Fig. 5(b) and Fig. 5(c). Both are able to significantly reduce the background noise. Both require empirical selection of the parameters to reduce the possibility of signal loss due to over subtraction and to limit the background and musical noise. During the non-speech segment, the HNM based spectral subtraction gives somewhat larger noise. However during the speech signal it appears to be more effective.

While synthesizing the speech using HNM, the noise part was not added, as the output was very intelligible and background noise free if synthesized using only harmonic part. So we do not require the processing of noise part while analysis and synthesis using HNM. It reduces the number of parameters along with the processing time required for synthesis. Speech processed by HNM contains less musical noise. Spectrographic analysis shows that the formants structure in HNM synthesized speech is more enhanced as compared to the original and that of processed with out HNM.

6. CONCLUSION

Our group has earlier reported pitch synchronous application of spectral subtraction for reduction of leakage noise in electrolaryngeal speech [10]. Here we have presented an investigation involving only the harmonic part of pitch synchronous harmonic plus noise model (HNM) of speech [14] for spectral subtraction. It is seen that this method provides an effective subtraction of noise during the speech segments, and hence can be used for improving intelligibility of electrolaryngeal speech.

In this technique, the noise spectrum is estimated during silence period, with speaker turning the vibrator on but keeping the lips closed. It has been observed that this noise estimate does not remain valid after a lapse of time, due to changes in the application pressure on the vibrator, and the estimate has to be updated frequently. Our group has investigated the use of quantile based noise estimation (QBNE) for continuously updating of the noise spectrum without involving speech / non-speech detection [12],[15]. QBNE can be combined with HNM based spectral subtraction presented in this paper.



a) Recorded speech signal

b) Processed by spectral subtraction using magnitude spectrum

c) Processed by spectral subtraction using HNM derived harmonic spectrum

Figure 5. Waveforms for recorded and processed speech. Speaker SP, material: question-answer pair in English, "*What is your name? My name is Santosh*".

7. REFERENCES

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

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

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

[4] Childers, D.G. and Hu, T.H. Speech synthesis by glottal excited linear prediction in J. Acoust Soc. Am., 96, 4, (1994), 2026-2036.

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

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

[7] Lehana, P.K. and Pandey, P.C. Speech synthesis in Indian languages in Proc. Int. Conf. on Universal Knowledge and Languages-2002, (Nov. 25-29, 2002), Goa, India, , paper no. pk 1510.

[8] McAulay, R.J. and Quatieri, T.F. Speech analysis/synthesis based on a sinusoidal representation in IEEE Trans. Acoust, Speech, Signal Processing, 34 (no. 4, 1986), 744-754.

[9] O'Saughnessy, D. Speech Communication - Human and Machine. Universities press, Hyderabad, 2001.

[10] Pandey, P.C., Bhatnagar, S.M., Bachher, G.K., Lehana, P.K. Enhancement of alaryngeal speech using spectral subtraction in DSP2002 (1-3 July 2002), Santorini, Greece, 591-594.

[11] 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).

[12] Pratapwar, S. S., and Pandey, P. C. Quantile based noise estimation for spectral subtraction of self leakage noise in electrolaryngeal speech in Proc. Int. conf. systemics, cybernetics, and informatics, ICSCI 04

[13] Rabiner, L.R. and Schafer, R.W. Digital Signal Processing of Speech Signals. Prentice Hall, Engelwood Cliffs, New Gersy: l, 1978.

[14] Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis in IEEE Trans. Speech Audio Processing, 9,1. (2001), 21-29.

[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.