Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Enhancement of electrolaryngeal speech by spectral subtraction, spectral compensation, and introduction of jitter and shimmer

Prem C. Pandey (1) and S. Khadar Basha (2)

(1, 2) Indian Institution of Technology Bombay, Powai Mumbai 400 076, Maharashtra, India

PACS: 43.72.AR

ABSTRACT

An electrolarynx, a verbal communication aid used by laryngectomy patients, is a vibrator held against the neck tissue to provide excitation to the vocal tract, as a substitute to that provided by the glottal vibrations. Although the user can set the vibration level and pitch, a dynamic control of level, voicing, and pitch during speech production is not feasible. In addition to this basic limitation, the electrolaryngeal speech suffers from (i) presence of background noise caused by leakage of acoustic energy from the vibrator and vibrator-tissue interface, (ii) low-frequency spectral deficiency, and (iii) unnatural quality due to constant pitch and level. Background noise decreases the intelligibility, while the other two factors affect the speech quality. Present study involved investigations for improving the intelligibility and quality of electrolaryngeal speech. Pitch-synchronous application of generalized spectral subtraction was used for reducing the background noise. In order to track the variation in the spectrum of the leakage noise due to changes in vibrator orientation and pressure during speech production, a dynamic estimation of noise was carried out from a set of past frames. The estimated noise spectrum was subtracted from that of the noisy speech and the resulting magnitude spectrum was combined with the original phase spectrum. The speech signal was resynthesized using overlap-add method, with two-pitch period analysis frames and one period overlap. Estimation of phase spectrum by minimum-phase assumption and the assumption of phase continuity did not improve the speech quality. An introduction of jitter and shimmer in the speech signal, using LPC based analysis-synthesis, was investigated for improving its naturalness. The excitation for synthesis was an impulse train with the frequency equal to that of the vibrator, with random frequency and amplitude modulations for providing the jitter and the shimmer, respectively. An FIR filtering of the excitation was used to match the long-term average spectral envelope of the processed electrolaryngeal speech to that of the normal speech. A peak-to-peak jitter of up to 6 % increased the naturalness, while introduction of shimmer decreased the quality.

INTRODUCTION

The artificial larynx is a verbal communication aid used by laryngectomy patients for providing a voicing source, as an alternative to the glottis in the natural larynx. Electrolarynx, or the external electronic larynx, is the most widely used type of artificial larynx. It is a battery powered hand-held electronic vibrator. A schematic of speech production using this device is shown in Figure 1. Pulses from its vibrating diaphragm, held against the throat, get transmitted through the neck tissue to the vocal tract. The resonances of the time-varying vocal tract filter dynamically shape the harmonic spectrum of the vibrations. The resulting speech is known as electrolaryngeal speech. The devices generally permit setting of the vibration level and pitch by the user. However, a dynamic control of level, voicing, and pitch during speech production is not feasible. In addition to this basic limitation, the electrolaryngeal speech suffers from (i) presence of background noise caused by leakage of acoustic energy from the vibrator and vibrator-tissue interface, (ii) low-frequency spectral deficiency due to attenuation of the lower harmonics in transmission through the neck tissue, and (iii) unnatural quality due to constant pitch and level. Background noise decreases the intelligibility, while the other two factors affect the speech quality [1]-[3].

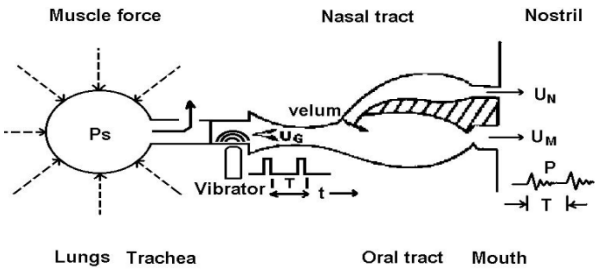

Figure 1. Speech production using an electrolarynx [6].

It has been reported that acoustic shielding of the vibrator assembly could reduce the leakage of the acoustic energy from the vibrator, but the shielding effect of the insulation was counterbalanced by mechanical damping and it was not effective in reducing the leakage from the vibrator-tissue interface [4]. Several speech-processing techniques have been reported for enhancing the electrolaryngeal speech [4]-[10]. Present study involved investigations for improving the intelligibility and quality of electrolaryngeal speech, with the objective of selecting the appropriate techniques for real-time implementation. Generalized spectral subtraction with a

dynamic estimation of the spectrum of the background noise without using a speech activity detector is used for suppressing the background noise. Effect of different methods of phase estimation is also investigated. A filter is used to approximately compensate for the spectral deficiencies. Finally, effect of introducing jitter and shimmer using LPC based analysis-synthesis for removing the mono-tonicity of the electrolaryngeal speech is also investigated.

SPECTRAL SUBTRACTION

In the spectral subtraction for enhancement of noisy speech, an estimate of the spectrum of the noise is subtracted from that of the noisy speech and the resulting magnitude spectrum is combined with the phase spectrum of the noisy speech for resynthesizing the clean speech [11], [12]. Several methods have been reported for dynamically estimating the noise spectrum and to take care of the short-term variations in the noise spectrum [13], [14]. The method is based on the assumption that the speech and the noise are uncorrelated. In electrolaryngeal speech, the speech signal and the background noise originate from the pulsatile vibrations of the diaphragm and hence they are strongly correlated. It has been shown in [5] that if the spectra are calculated pitch-synchronously, the speech and noise become uncorrelated and spectral subtraction can be employed.

A model of the generation of the background leakage noise in electrolaryngeal speech is shown in Figure 2. The impulse response of the vocal tract filter and the impulse response of the leakage path are represented as $h_v(n)$ and $h_l(n)$, respectively. The speech signal $s(n)$ and the leakage noise $l(n)$ are generated by convolution of the pulsatile excitation $e(n)$ with the respective impulse responses

$$s(n) = e(n) * h_v(n) \quad (1)$$

$$l(n) = e(n) * h_l(n) \quad (2)$$

The noisy speech signal is given as

$$x(n) = s(n) + l(n) \quad (3)$$

The vocal tract acts as a time-varying filter during speech production, while the filter response of the leakage path varies slowly due to changes in the orientation and pressure in holding the vibrator against the neck tissue. Applying short-time Fourier transform on (3), we get

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

The impulse responses of the vocal tract filter and the leakage path may be assumed to be uncorrelated, and hence

$$|X_n(e^{j\omega})|^2 = |E_n(e^{j\omega})|^2 [|H_{vn}(e^{j\omega})|^2 + |H_{ln}(e^{j\omega})|^2] \quad (5)$$

If a pitch-synchronous window is used to evaluate short-time spectra, $|E_n(e^{j\omega})|^2$ may be considered as constant $|E(e^{j\omega})|^2$. During the non-speech intervals, $s(n)$ will be negligible and the noise spectrum is given as

$$|L_n(e^{j\omega})|^2 = |E(e^{j\omega})|^2 |H_{ln}(e^{j\omega})|^2 \quad (6)$$

The noise spectrum can be estimated from the noise during explicit silences (lips closed), or dynamically using a voice activity detector. It can also be estimated using statistical techniques without a voice activity detector.

A block diagram of the spectral subtraction technique is shown in Figure 3. All the spectral estimates are computed

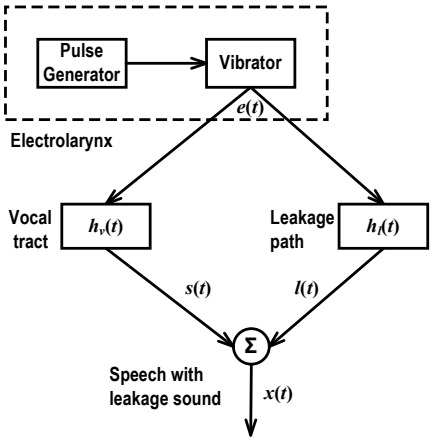

Figure 2. A model of the background leakage noise generation in elecctrolaryngeal speech [5].

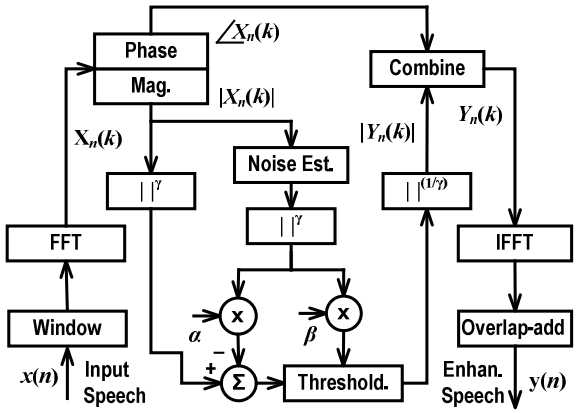

Figure 3. Block diagram of spectral subtraction.

using FFT. In the generalized spectral subtraction technique [12], the cleaned magnitude spectrum is obtained as

$$E(k) = |X_n(k)|^\gamma - \alpha|L_n(k)|^\gamma \quad (7)$$

$$|Y_n(k)| = [E(k)]^{(1/\gamma)}, \text{ if } E(k) > [\beta|L_n(k)|]^\gamma$$

$$\beta|L_n(k)|, \text{ otherwise} \quad (8)$$

where α is an oversubtraction factor used to reduce the residual noise due to short-time variations in the noise. Oversubtraction may result in negative values in the spectrum, causing time-varying tonal sounds, known as "musical noise", which adversely affect the quality of the resynthesized speech. This noise is masked by a floor noise, controlled by the floor factor β. The subtraction power γ = 2 results in power subtraction and γ = 1 results in magnitude subtraction. The values of the three parameters need to be empirically obtained for each type of noise estimation and the device. The magnitude spectra after spectral subtraction are combined with the corresponding phase spectra of the noisy speech and the resulting complex spectra are used to resynthesize speech by using overlap-add method.

In addition to the resynthesis using the original noisy phase, effect of estimating the phase spectrum by other methods was also investigated: (i) zero phase, (ii) randomly selected phase, (iii) phase set for continuity across the frames, and (iv) phase spectrum estimated from the spectrally subtracted magnitude

spectrum using the assumption of minimum-phase signal. The minimum-phase estimation was carried out using iterative technique [15] and cepstrum-based non-iterative technique [16]-[18].

ESTIMATION OF NOISE SPECTRUM

The characteristics of the background noise due to leakage of acoustic energy from the vibrator are generally different from those of other kinds of background noise. Its spectrum slowly varies due to the changes in the orientation of the electrolarynx against the neck tissue and the hand pressure in holding it during speech production. Estimation of the noise during silence intervals of speech needs a voice activity detector, but it is difficult to reliably separate the voice and silence segments in electrolaryngeal speech. Pandey *et al.* [5] used an averaging based noise estimation during the initial 2 s silence period of a recording, but this method is not suitable for long-term use. Use of a quantile-based noise estimation [13] without an explicit voice/silence detector was reported to be effective in tracking the electrolaryngeal noise. In this method the quantile values for different spectral components were selected for matching the noise spectrum estimated over a long speech record to match that obtained by averaging during the initial silence [6]. The method is difficult to implement for real-time processing. Liu *et al.* [7], [8] reported spectral subtraction with adaptation of parameters using frequency domain masking properties of the auditory system for suppression of the leakage noise as well as the external noise. Mitra and Pandey [9] and Kabir *et al.* [10] used the minimum statistics based method of Martin [14] for dynamically estimating the noise without speech-nonspeech discrimination. This method is computationally less expensive and is suitable for real-time implementation if the subtraction parameters do not have to be dynamically estimated from the signal statistics.

We investigated the use of averaging based noise estimation, median based noise estimation (a simple case of quantile based noise estimation) and minimum statistics based estimation for tracking the noise in electrolaryngeal speech, by using fixed values of subtraction parameters and frames.

INTRODUCTION OF JITTER AND SHIMMER AND SPECTRAL COMPENSATION

Random variations in the level and the pitch in speech are known as the jitter and the shimmer, respectively. Electrolaryngeal speech sounds monotonous and unnatural, as it has no jitter and shimmer. While a dynamic control of voicing, pitch, and level by the user of the device is not feasible, introduction of jitter and shimmer in the electro-laryngeal speech, either by introducing it in the vibrator itself or by processing of the signal after suppression of the background noise, may help in reducing its unnaturalness.

For investigating the effect of jitter and shimmer in electrolaryngeal speech, a LPC based analysis-synthesis, as shown in Figure 4, was used. The time-varying response of the vocal-tract filter was estimated by LPC analysis [18] and the coefficients of the prediction filter were used to realize a time-varying filter for resynthesizing the speech. The LPC analysis was carried out using 2-pitch period window and autocorrelation method for estimating 12 predictor coefficients. To closely track the vocal tract variation, 5-sample frame shifting was used. The time-varying resynthesis filter was excited by an impulse train with its frequency equal to that of the vibrator. Shimmer is introduced by varying the amplitude of the impulses as $a(1+sr_1)$, where a is the amplitude, r_1 is a random number uniformly distributed over

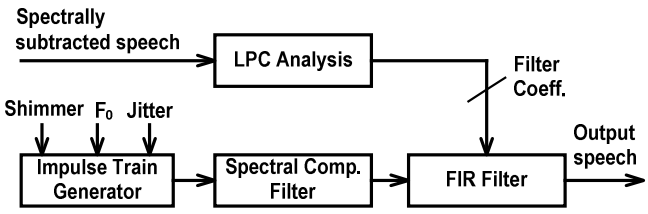

Figure 4. Introduction of jitter, shimmer, and spectral compensation using LPC based analysis-synthesis.

± 0.5, and s is the peak- to-peak shimmer. Jitter is introduced by varying the spacing of the successive impulses as $N(1+jr_2)$, where N is the pitch period in number of samples, r_2 is a random number uniformly distributed over ± 0.5, and j is the peak-to-peak jitter.

Electrolaryngeal speech is deficient in low frequency content due to a relatively higher attenuation of low frequency components during the transmission of the vibrations through the neck tissue. Use of an impulse train as the excitation in the LPC based analysis-synthesis resulted in an emphasis of high frequency in the resynthesized speech. A spectral compensation filter was inserted in the excitation path to approximate the long-duration averaged spectrum of the resynthesized signal to that of the natural speech. Sustained vowels /a/, /i/, /u/ were recorded from a speaker speaking naturally and by using an electrolarynx. Ratio of the averaged LPC-smoothened spectra of the natural speech and the electrolaryngeal speech after spectral subtraction was used to obtain the magnitude spectrum of the compensation filter and the filter was designed as a linear-phase FIR filter.

RESULTS AND DISCUSSION

Electrolaryngeal speech was recorded from two normal speakers, using electrolarynx models SolaTone (pitch frequency = 126.7 Hz) and NP-Voice (93.4 Hz), at a sampling rate of 11.025 kHz and 16-bit quantization. Spectral subtraction was performed using 2-pitch period frames with 50 % overlap. All the processing was carried out using Matlab. Effects of spectral subtraction, frequency compensation, and introduction of jitter and shimmer were assessed through informal listening tests.

The optimal values of the three factors in the generalized spectral subtraction were found to be dependent on the noise estimation method. It was found that use of power $\gamma = 1$ resulted in more tolerance to the variations in the values of the over-subtraction factor α and the floor factor β. For the noise estimated by averaging the noise during initial 2-s segment with lips closed, best results were obtained with $\alpha = 2.0$ and $\beta = 0.001$. However, the noise estimation was effective for spectral subtraction only up to about 5 s. With minimum statistics based noise estimation, best results were obtained for $\alpha = 5.0$ and $\beta = 0.005$. The method was found to need about 4 s of silence for correctly estimating the noise. It was found that in the absence of frequent pauses in speech, the noise estimation was affected by speech segments and resulted in distortion of speech. Median based noise estimation was able to track the noise without requiring a long initial silence or frequent pauses in speech. Best results were obtained with $\alpha = 1.2, \beta = 0.001$.

Investigations with different estimations of the phase spectrum showed that the speech quality for both the types of minimum-phase estimation were similar and not better than that obtained by using the phase of the noisy speech. Use of zero and random phases resulted in poor quality.

An example of noise suppression is shown using the waveforms and spectrograms in Figure 5, for the original electrolaryngeal speech, and the speech after spectral subtraction, resynthesis by LPC-based analysis-synthesis, and spectral compensation.

Use of compensation filter significantly improved the quality of the speech. Speech was resynthesized by introducing jitter and shimmer with the peak-to-peak values varied from 0 to 40 %. A peak-to-peak jitter of 6 % resulted in maximum improvement in naturalness, while the values above 20 % resulted in degradation of speech. Introduction of shimmer up to 20 % did not result in an improvement in naturalness, while the larger values of shimmer degraded the speech.

CONCLUSION

The investigations showed that the magnitude spectral subtraction using median-based noise estimation and resynthesis using noisy phase was effective in suppression of background noise in electrolaryngeal speech. While introduction of shimmer did not help, introduction of peak-to-peak jitter of 6 % and spectral compensation further increased the quality of the resynthesized speech. Listening tests on speech recorded from a number of laryngectomy patients need to be conducted for a detailed evaluation of the intelligibility and quality of the processed speech.

ACKNOWLEDGMENT

The research is partly supported by a project grant under the National Programme on Perception Engineering, sponsored by the Department of Information Technology, MCIT, Government of India.

REFERENCES

1 M. Weiss, G. Y. Komshian, and J. Heinz, "Acoustic and perceptual characteristics of speech produced with an electronic artificial larynx," *J. Acoust. Soc. Am.,* **65,** 1298-1308 (1979).
2 H. L. Barney, F. E. Haworth, and H. K. Dunn, "An experimental transistorized artificial larynx," *Bell Systems Tech. J.,* **38**, 1337-1356 (1959).
3 Q. Yingyong and B. Weinberg, "Low frequency energy deficit in electrolaryngeal speech," *J. Speech Hearing Res.*, **34**, 1250-1256 (1991).
4 C. Y. Espy-Wilson, V. R. Chari, and C. B. Haung, "Enhancement of alaryngeal speech by adaptive filtering," *Proc. ICSLP,* 764-771 (1996).
5 P. C. Pandey, S. M. Bhandarkar, G. K. Baccher, and P. K. Lehena, "Enhancement of alaryngeal speech using spectral subtraction," *Proc. 14th Int. Conf. Digital Signal Prcessing (DSP 2002),* Santorini, Greece, 591-594 (2002).
6 P. C. Pandey, S. S. Pratapwar, and P. K. Lehana, "Enhancement of electrolaryngeal speech by reducing leakage noise using spectral subtraction with quantile based dynamic estimation of noise," *Proc. 18th Int. Congress Acoustics (ICA 2004),* Kyoto, Japan, 3029-3032 (2004).
7 H. Liu, Q. Zhao, M. Wan, and S. Wang, "Application of spectral subtraction method on enhancement of electrolaryngeal speech," *J. Acoust. Soc. Am.*, **120**, 398-406 (2006).
8 H. Liu, Q. Zhao, M. Wan and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomed. Eng*., **53**, 865-874 (2006).
9 P. Mitra and P.C. Pandey, "Enhancement of electro-laryngeal speech by spectral subtraction with minimum statistics-based noise estimation," *J. Acoust. Soc. Amer.*, **120**, 3039 (2006).
10 R. Kabir, A. Greenblatt, K. Panetta, and S. Agaian, "Enhancement of alaryngeal speech utilizing spectral subtract ion and minimum statistics," *Proc. 7th International Conference on Machine Learning and Cybernetics,* Kunming, 12-15 July (2008).
11 S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process,* **27**, 113-120 (1979).
12 M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc.IEEE ICASSP'79*, 208-211 (1979).
13 V. Stahl, A. Fisher, and R. Bipus, "Quantile based noise estimation for spectral subtraction and wiener filtering," *Proc. IEEE ICASSP'00*, **3**, 1875-1878 (2000).
14 R. Martin, "Spectral subtraction based on minimum statisic" *Proc. 7th European Signal Processing Conf. (EUSIPCO–94),* Edinburgh, Scoltland, 1182-1185 (1994).
15 T. F. Quatieri and A. V. Oppenheim,"Iterative techniques for minimum phase signal reconstruction from phase or magnitude," *IEEE Trans. Acoust., Speech, Signal Process.,* **29**, 1187-1193 (1981).
16 B. Yegnanarayana and A. Dhayalan, "Noniterative techniques for minimum phase signal reconstruction from phase or magnitude," *Proc. IEEE ICASSP,* 639-642, (1983).
17 A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. (Prentice-Hall, Englewood Cliffs, New Jersey, 1975).
18 L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* (Prentice Hall, Englewood Cliffs, New Jersey, 1978).

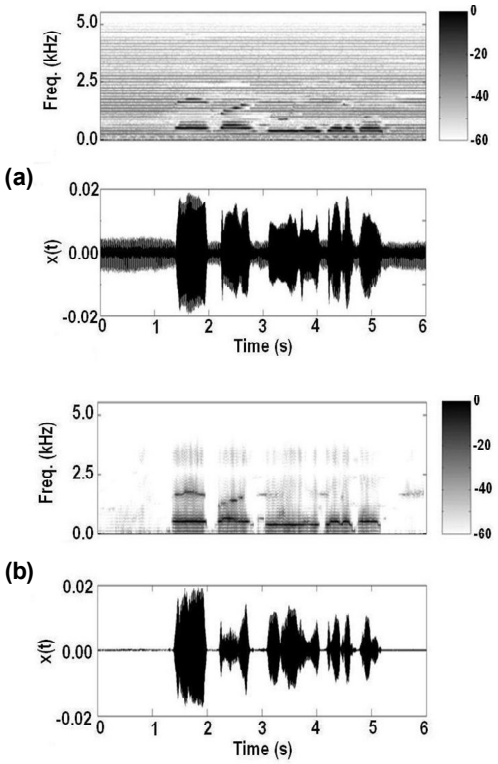

Figure 5. Example of processing: (a) Electrolaryngeal speech for the sentence ' … *Where were you a year ago ?',* (b) Processed output using median based noise estimation with 400 frames, α =1.2, β =0.001, γ = 1, j = 0.06, and s = 0.