MATTER: International Journal of Science and Technology
ISSN 2454-5880

Akdeniz & Oktay, 2017

Volume 3 Issue 2, pp. 433-444

Date of Publication: 7th November 2017

DOI-https://dx.doi.org/10.20319/mijst.2017.32.433444

This paper can be cited as: Akdeniz, R., & Oktay, M. (2017). Comparison for Speech Coding Algorithms for Total Laryngectomies. MATTER: International Journal of Science and Technology, 3(2), 433-444.

COMPARISON FOR SPEECH CODING ALGORITHMS FOR TOTAL LARYNGECTOMIES

Rafet AKDENİZ
Department of Electronics and Telecommunication Engineering, Çorlu Engineering Faculty, Namık Kemal University, 59860 Çorlu, Tekirdağ, Turkey

Mesut Oytun OKTAY
MTC Makina ve Otomasyon, Veliköy Organize Sanayi Bölgesi, Mimar Sinan Mah. 47. Sok. No:659500 Karaağaç, Çerkezköy Tekirdağ, Turkey

Abstract

Electrolarynx is used as a noninvasive supporting device for speech restoration in people who have undergone resection operation over their larynxes. This work aims to develop a signal processing method to neutralize the mechanical vibration noise of this device. We investigate the effect of this noise on the speech signal and analyze the performances of various algorithms in a single input system to minimize this noise.

Keywords

Total Laryngectomy, Speech Enhancement, Amplitude Spectral Subtraction, Power Spectral Subtraction

1. Introduction

This study will scrutinize the voice distortion witnessed in people who had their larynxes removed (who had a total laryngectomy) due to throat cancer, and the improvement of the electrolarynx speech, one of the voice rehabilitation methods of post total laryngectomy will be analyzed.

Throat cancer is the most common malign tumor witnessed in the head and neck (Genden at al., 2007). When radiotherapy and protective throat surgery (conservative laryngeal surgery) fail to address the situation, partial or total laryngectomy, the removal of the larynx, can be used as a method (Aksoy, Veyseller, Yıldırım, Demirhan & Özturan, 2010).

After total laryngectomy, the vibrator organ, one of the three organs that form the voice: activator, vibrator and articulator, is lost (Erişir & İnci, 2001). After the operation, a way is sought to give the patient back his voice. Nowadays the paths that are taken most frequently are esophagus speech, speech aided by electro larynx (electrolarynx) and operational voice restoration. Whichever voice restoration used after the total laryngectomy, this new voice is different than the patient's original voice and the natural human voice.

With digital signal processing becoming more common, improving this voice is now a common area of study and a commonly tackled problem. The main differences of this restored speech compared to the normal speech are the lower and constantly changing main frequency, lower volume, and a shift towards higher pitch due to the shortening of the voice way in formant frequencies (Cole, Sridharan, Moody & Geva, 1997 Dec.).

Efforts at speech enhancement are generally focused on these kind of dysphonic speeches. The main approach for improving these is to stimulate the formant acquired through linear predictive coding by simulated glottal wave structure. Exemplary works can be seen in (Türkmen, 2008) and (Tarakçıoğlu. 2010). This way of improving has been seen to result better than formant frequency stimulations (Cole et al., 1997 Dec.).

Electronic larynx is a device at the size of palm, placed under the neck or inside the mouth in order to create the vibration necessary for speaking. However, it is not preferred because it occupies one hand when used, produces a rather mechanical voice, and creates dependency on a battery and its high price (Aksoy et al., 2010). On the other hand, its advantages are that it requires little effort in long sentences and it is effective in many circumstances (Liu, Zhao, Wan & Wang 2006).

The main difference between electronic larynx speech and regular speech is the former having always the same main frequency because it is mechanically produced. It is also distinguished from other speeches produced via different methods with this quality. In addition, this feature makes it impossible as of today to create the required stops when forming unvoiced phonemes. Finally, the basic frequency changes necessary to transmit emotions and emphases cannot be done with this device (Niu, Wan, Wang & Liu, 2003). This method's speech enhancement efforts are focused on suppressing the device's mechanical voice.

Two main approaches are followed in suppressing the signal level generated by the device. These approaches involve the usage of single and multi-sensors, and in both cases, the correct detection of the speech is important. In the methods involving multiple sensors, thanks to the sensor in which the noise is the main component, noise component is acquired and extracted from the signal where noise is the passive component. In the methods involving single sensor, noise prediction is made from the speechless moments.

The intervention level depends on the device as well as on the patient. This difference is measured to be between 7- 26 dB in the measurements done on 30 patients, while their mouths were shut (Niu et al., 2003). These differences are attributed to the differences in how the device is coupled to the neck, the characteristics of vibration, the propagation characteristics of the neck, and the user's competence. This study evaluates natural voice acquisition competences of the single sensor methods of noise subtraction.

2. Methods

It is thought that what makes the electronic larynx speech unnatural is the noise dissipated from the device (Meltzner, 2003). Therefore, focus was on the algorithms of noise subtraction to suppress this noise in the single sensor system. The path taken is to collect data from subjects, preprocessing this data and to process it with determined algorithms, then sending it to a group without subjects, in order to assess the success of the algorithms.

2.1. Data Gathering

Subjects are chosen among healthy individuals and were expected to get used to the electro larynx in a short amount of time. Of the two young adults between the ages of 20 and 30, one man and woman, the woman could not adapt to the electronic larynx.

Sample sentence is *"Bize reality show izlettireceğiz diye, ortalığı salhaneye çevirmeye, mezbahaya döndürmeye hiç gerek yok"* in Turkish was asked to be read by those who could use the electronic larynx was made to repeat three times, giving pauses in between. The recording took place in a large and silent room, and was conducted with an *M-Audio Fast Track external sound card* was connected to a *Samson C01 Studio Condenser microphone*, which was hold about 20 cm away from the subjects' mouth. Recordings are conducted via Audacity voice recording and editing software, and sampling rate is set as 48 KHz.

2.2. Preprocessing

Speeches recorded are first passed through a high pass filter of 50 Hz, then a low pass filter of 9,6 KHz. Filtered signal is divided into windows of 30 ms that overlap with an amount of 75%, and Hamming is used as windowing function.

2.3. Noise Prediction

Assume that y(n), the noise-corrupted input signal, is composed of the clean speech signal *x(n)* and the additive noise signal, *d(n)*, that is (Loizou, 2007);

$$y(n) = x(n) + d(n) \tag{1}$$

Taking the discrete-time Fourier transform (DTFT) of both sides gives

$$Y(\omega) = X(\omega) + D(\omega) \tag{2}$$

We can express $Y(\omega)$ in polar form as follows:

$$Y(\omega) = |Y(\omega)| e^{j\phi_y(\omega)} \tag{3}$$

Where, $Y(\omega)$ is the magnitude spectrum and $\phi_y(\omega)$ is the phase of the corrupted noisy signal.

The noise spectrum $D(\omega)$ can also be expressed in terms of its magnitude and phase spectra as $D(\omega) = |D(\omega)| e^{j\phi_d(\omega)}$. The magnitude noise spectrum is unknown, but can be replaced by its average value computed during non-speech activity. Similarly, the noise phase $\phi_d(\omega)$ can be replaced by the noisy speech phase $\phi_y(\omega)$. This is partly motivated by the fact that phase does not affect speech intelligibility. After then we can obtain an estimated of the clean speech signal spectrum:

$$\hat{X}(\omega) = [|Y(\omega)| - |\hat{D}(\omega)|] e^{j\phi_y(\omega)} \tag{4}$$

Where, $\left|\hat{D}(\omega)\right|$ is the estimated of the magnitude noise spectrum that made during non-speech activity. The symbol "^" indicates estimated parameters.

2.4. Processing

Speech enhancement is done as the following: in moments marked as noise, noise prediction was done. While in other moments, predicted noise is subtracted via defined methods. For comparison, this study used six different methods. These are;

- Amplitude spectral subtraction,
- Power spectral subtraction,
- Spectral subtraction using over-subtraction,
- Non-linear spectral subtraction,
- Multiple band spectral subtraction,
- Spectral subtraction using auditory masking.

2.4.1. Amplitude Spectral Subtraction (ASP)

This method involves subtracting the noise's amplitude D from the signal's amplitude X in the frequency space. During the subtraction, spectral amplitude components that fall into the negative due to erroneous noise prediction are moved to zero (Loizou, 2007), (Berouti, Schwartz & Makhoul, 1979 April). The method's mathematical expression in terms of window is shown in (5).

$$|\hat{X}(\omega)| = \begin{cases} |Y(\omega)| - |\hat{D}(\omega)| & if\ |Y(\omega)| > |\hat{D}(\omega)| \\ 0 & else \end{cases} \tag{5}$$

2.4.2. Power Spectral Subtraction (PSS):

This method involved subtracting the predicted noise's spectral power components from the signal's spectral power components (Loizou, 2007). The method's mathematical expression is shown in (6).

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - |\hat{D}(\omega)|^2 & if\ |Y(\omega)|^2 > |\hat{D}(\omega)|^2 \\ 0 & else \end{cases} \tag{6}$$

Rectifying was used in this method, too.

2.4.3. Spectral Subtraction Using Over-Subtraction (SSUOV)

The Amplitude spectral subtraction and Power spectral subtraction methods are not concerned with the acquired spectrum. They involve the subtraction of predicted noise's spectral amplitude/power from the speech signal's spectral amplitude/power. Only, rectifying was used so as to ensure that the acquired spectrum is consistent.

Because of the errors made in noise prediction, the rectifying causes isolated hills to form in the amplitude spectrum. Those isolated hills in certain frequencies form voices similar to squeaking. This noise is known as musical noise.

To remove the isolated hills, over-subtraction method is offered in (Loizou, 2007). This method involves subtracting predicted noise from the processed voice, thus removing all isolated hills in narrow or wide bands. To lessen the effect of the hills that remain after excessive inference, part of the subtracted noise is added back to the improved signal. The method's mathematical expression is shown in (7).

$$|\hat{X}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha|\hat{D}(\omega)|^2 & if\ |Y(\omega)|^2 > (\alpha+\beta)|\hat{D}(\omega)|^2 \\ \beta|\hat{D}(\omega)| & else \end{cases} \quad (7)$$

In these expressions, the α is the over-subtraction multiplier, while β is the spectral base parameter. α is chosen as $\alpha > 1$ and close to 1, whereas β is chosen as $0 < \beta << 1$ (Loizou, 2007). The multipliers α and β used in this method change depending on the signal to noise ratio (SNR), calculated by the expression given in (8). The expressions are as follows:

$$\alpha = \begin{cases} \alpha_0 - \frac{-5}{s} & if \quad SNR \leq -5dB \\ \alpha_0 - \frac{\mathrm{SNR}}{s} & if\ -5dB < SNR < 20dB \\ 1 & if \quad 20dB \geq SNR \end{cases} \quad (8)$$

$$\beta = \begin{cases} 0.02 & if \quad 0dB < SNR \\ 0.06 & if \quad 0dB \geq SNR \end{cases} \quad (9)$$

In the expression for calculating the over-subtraction multiplier, s is the value that ensures α =1 in the upper limit of SNR. In the application, for α_0 = 3dB, s = 20/2. This slope is experimentally specified (Loizou, 2007).

2.4.4. Non-Linear Spectral Subtraction (NLSS)

This is an approach designed taking into account that noise does not affect signal the same way for each component in the spectrum. Non-linear spectral subtraction can be defined as the over-subtraction multiplier of the spectral subtraction where over-subtraction is used, taking on different values changing with frequency. Non-linear spectral subtraction rule is shown as follows in (10).

$$|\hat{X}(\omega)| = \begin{cases} |\bar{Y}(\omega)| - \alpha(\omega)N(\omega) & if\ |\bar{Y}(\omega)| > \alpha(\omega)N(\omega) + \beta.|\bar{D}(\omega)| \\ \beta|\bar{X}(\omega) & else \end{cases} \quad (10)$$

In the equation (10), β is spectral floor, $|\bar{Y}(\omega)|$ and $|\bar{D}(\omega)|$ are the smoothed estimates of noisy speech and noise, respectively. $\alpha(\omega)$ is a frequency-dependent subtraction factor, $N(\omega)$ is a nonlinear function of the noise spectrum. The smoothed estimates of noisy speech $|\bar{Y}(\omega)|$ and noise $|\bar{D}(\omega)|$ are obtained as follows:

$$|\bar{Y}(\omega)_i| = \mu_y |\bar{Y}_{i-1}(\omega)| + (1-\mu_y)|\bar{Y}_i(\omega)| \quad (11)$$

$$|\bar{D}_i(\omega)| = \mu_d |\bar{D}_{i-1}(\omega)| + (1-\mu_d)|\hat{D}_i(\omega)| \quad (12)$$

Where, the constants, μ_d take values in range $0.1 \le \mu_y \le 0.5$ 0.5 and $0.5 \le \mu_d \le 0.9$ [10]. In the working algorithm, both smoothing coefficients are chosen as 0.5.

In the expression (12), the *N(ω)* term is obtained by computing the maximum of the noise magnitude spectra, $|\hat{D}_i(\omega)|$, over the past 40 frames.

$\alpha(\omega)$ is constant for all frequencies but varies from frame to frame depending on the a posteriori SNR. For the $\alpha(\omega)$ in equation (10), which forms the basis of the method is given as follows:

$$\alpha(\omega) = \frac{1}{1+\gamma\rho(\omega)}, \qquad \rho(\omega) = \frac{|\bar{Y}(\omega)|}{|\bar{D}(\omega)|} \quad (13)$$

Where γ is a scaling factor, and $\rho(\omega)$ is the square root of the posteriori SNR estimate, which defines the impact of the signal to noise ratio in the algorithm is considered ρ=1.

2.4.5. Multiband Spectral Subtraction (MSS)

This method is put forward to make the non-linear spectral inference stronger and resistant against the potential distortion caused by sudden changes in the noise. It does this by defining the inference coefficient on a band basis (Loizou, 2007).

The noiseless signal's expression is shown , the band being *i*, band beginning frequency being b_i and the band ending frequency being e_i, as follows:

$$|\hat{X}_i(\omega_k)|^2 = |\bar{Y}_i(\omega_k)|^2 - \alpha_i \delta_i |\hat{D}_i(\omega_k)|^2 \quad b_i \le \omega_k \le e_i \tag{14}$$

Where, $\omega_k = 2\pi k / N \ (k = 0,1,...,N-1)$ are the discrete frequencies, $|\hat{D}_i(\omega_k)|^2$ is the estimated noise power spectrum, b_i and e_i are the beginning and ending frequency bins of the ith frequency band, α_i is the over subtraction factor of the ith band, δ_i is an additional band-subtraction factor that can be individually set for each frequency band to customize the noise removal process.

A weighted spectral average is taken over preceding and succeeding frames of speech as follows:

$$|\bar{Y}_j(\omega_k)| = \sum_{k=-M}^{M} W_i |Y_{j-i}(\omega_k)| \tag{15}$$

Where, $|\bar{Y}_j(\omega_k)|$ and $|Y_{j-i}(\omega_k)|$ are the preprocessed noisy magnitude spectrum of the jth frame and the noisy magnitude spectrum respectively. The weights W_i here are defined as (0.09, 0.25, 0.32, 0.25, 0.09) through experiments (Loizou, 2007). For each band, SNR is calculated separately:

$$SNR_i(dB) = 10log_{10}\left(\frac{\sum_{\omega_k = b_i}^{e_i} |\hat{Y}_i(\omega_k)|^2}{\sum_{\omega_k = b_i}^{e_i} |\hat{D}_i(\omega_k)|^2} \right) \tag{16}$$

And the over-subtraction coefficient on a band basis is calculated with the following formula:

$$\alpha_i = \begin{cases} 4.75 & SNR_i < -5 \\ 4 - \frac{3}{20} SNR_i & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \tag{17}$$

The δ_i coefficients that provide further control on the basis of frequency is shown as;

$$\delta_i = \begin{cases} 1 & f_i \le 1kHz \\ 2.5 & 1kHz \le f_i \le \frac{F_s}{2} - 2kHz \\ 1.5 & f_i > \frac{F_s}{2} - 2kHz \end{cases} \tag{18}$$

These values are acquired through experiments, as well (Loizou, 2007).

After these values are defined on the basis of window and band, basic subtraction equation (14) is applied to them, and to prevent amplitudes of smaller than zero appearing, the following is applied:

$$|\hat{X}_i(\omega_k)|^2 = \begin{cases} |\hat{X}_i(\omega_k)|^2 & if\ |\hat{X}_i(\omega_k)|^2 > \beta|\hat{Y}_i(\omega_k)|^2 \\ \beta|\bar{Y}_i(\omega_k)|^2 & else \end{cases} \tag{19}$$

In this formula, it is considered that β = 0.002. To suppress the musical noise, noise is added to the improved spectrum, and thus the improved signal is obtained:

$$|\bar{\bar{X}}_i(\omega_k)|^2 = |\hat{X}_i(\omega_k)|^2 + 0.05|\bar{Y}_i(\omega_k)|^2 \tag{20}$$

2.4.5. Perception Spectral Subtraction Through Auditory Mask (PSSTAM)

Auditory mask is related to people’s perception. Close to the high energy components of the signal, higher levels of noise can be perceived, whereas close to the signal’s low energy components, even lower levels of noise can be perceived (Virag, 1999). In this method, inference is made according to this threshold when noise is subtracted from the signal.

There are many ways to define this noise threshold. This study used the threshold definition method of using format frequencies (Liu et al., (2006).

$$P(z) = \frac{1 - \sum_{k=1}^{p} a_k \sigma_1^k z^{-k}}{1 - \sum_{k=1}^{p} a_k \sigma_2^k z^{-k}} \tag{21}$$

The perception threshold is considered σ_1 = 1, σ_2 = 0.8, in line with the inequation 0≤ σ_2 ≤ σ_1≤ 1 (Tarakçıoğlu, 2010). The a’s in the formula are coefficients of linear prediction coding.

After defining the perception threshold, the inference coefficient and the spectral base coefficient are defined, relying on the transformation $T(\omega) = P(e^{j\omega})$.

$$\alpha(\omega) = \alpha_{max}\left(\frac{T_{max}(\omega) - T(\omega)}{T_{max}(\omega) - T_{min}(\omega)}\right) + \alpha_{min}\left(\frac{T(\omega) - T_{min}(\omega)}{T_{max}(\omega) - T_{min}(\omega)}\right) \tag{22}$$

$$\beta(\omega) = \beta_{max}\left(\frac{T_{max}(\omega) - T(\omega)}{T_{max}(\omega) - T_{min}(\omega)}\right) + \beta_{min}\left(\frac{T(\omega) - T_{min}(\omega)}{T_{max}(\omega) - T_{min}(\omega)}\right) \tag{23}$$

Here, the following values are taken into consideration (Tarakçıoğlu, 2010):

$$\alpha_{max} = 6,\ \alpha_{min} = 1,\ \beta_{min} = 0 \text{ and } \beta_{max} = 0.02$$

2.5. Transition from a frequency space to time space

The output of all the methods is the improved amplitude spectrum. When going from here to the time space, the phase information of the noisy signal is used. The imaginary components that appeared upon the inverse Fourier Transformation are ignored.

3. Results And Conclusions

3.1. Results

Mean Opinion Score (MOS) is a subjective test which uses human opinion to assess the quality of the telephone network and nowadays is used for assessing the quality of signal processing algorithms. Mean opinion scoring is preferred in this work but applied with non-experts (Oktay & Akdeniz, 2014 April). To increase liability of the results, group of twenty four people is divided into two subgroups of tenant fourteen respectively. Outputs of the algorithms are given to the first group as it is and asked them to rate their quality from 1 (very bad) to 5 (very good). To the second group, outputs of the algorithms are doubled and shuffled randomly and given to the raters asked to score with the same scale. By doing this, we aim to measure the consistency of a score given to an algorithm by the rater. The value of a score is measured as the reciprocal of the variance of the scores given to the two sound files generated with the same algorithm. Weighted average of the MOS values are given in Table 1.

Table 1: *Mean Opinion Scores*

Algorithms	MOS [13]	MOS
ASS	2.643	2.393
PSS	3.786	2.536
SSUOS	1.643	2.321
NLSS	2.786	1.964
MSS	2.643	2.107
PSSTAM	2.143	2.393

Unprocessed speech	1.929	2.322

3.2. Conclusions

Our evaluation exhibited that, power subtraction was scored as the best algorithm. The fact that unprocessed voice scored worse than all but the excessive subtraction method scores to the fact that processing the input improves it to some degree. That all the scores given to the outputs have higher variation than the scores given to the input show that the group considers the input's intelligibility worse than the methods' outputs, however they are not in agreement over how good are the outputs.

Power subtraction method scored significantly higher than other methods. Amplitude and power subtraction methods can be termed as "parameter-less" methods among others. The success of these two methods that do not involve any parameters in their application can be attributed to the type of the noise. The other methods are the improved versions of the amplitude and power subtraction methods. Their lower scores, despite this fact, can be attributed to their parameters, which were not set according to this particular problem.

4. Acknowledgement

This work is supported by Namık Kemal University with project number NKUBAP.00.17.YL.11.01.

References

Aksoy F., Veyseller B., Yıldırım Y.S., Demirhan H. & Özturan O. (2010). Total larenjektomi sonrası ses restorasyonunda provox konuşma protezi tecrübemiz. Acta Oncologica Turcica. (43), pp. 65–69. (in Turkish)

Berouti M., Schwartz M. & Makhoul J. (1979 April). Enhancement of speech corrupted by acoustic noise. IEEE Int. Conf. Acoust. Speech Signal Processing. Washington DC, USA. pp. 208–211.

Cole D., Sridharan S., Moody M. & Geva S. (1997 December). Application of noise reduction techniques for Alaryngeal speech enhancement. IEEE TENCON'97 Speech and Image Technologies for Computing and Telecommunications. Queensland University of Technology, Brisbane, Australia. New York, NY, USA. pp. 491-495.

Erişir F. & İnci E. (2001). Total larenjektomiden sonra vocal rehabilitasyon. Cerrahpaşa Tıp Dergisi. 32 (2), pp. 80–85. (in Turkish)

Genden E.M., Ferlito A, Silver C.E., Jacobson A.S., Werner J.A., Suárez C., Leemans C.R., Bradley P.J. & Rinaldo A. (2007). Evolution of the management of laryngeal cancer. Oral Oncology, 43 (5), pp. 431–439. https://doi.org/10.1016/j.oraloncology.2006.08.007

Liu H., Zhao Q., Wan M. & Wang S. (2006). Enhancement of electrolarynx speech based on auditory masking. IEEE Transactions on Biomedical Engineering. 53 (5), pp. 864–875.

Loizou P.C. (2007). Speech Enhancement, Theory and Practice. Taylor- Francis, London, GB.

Meltzner G.S. (2003). Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech. PhD Thesis, Massachusetts Institute of Technology, Boston, USA.

Niu H.J., Wan M.X., Wang S.P. & Liu H.J. (2003). Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis. Medical and Biological Engineering and Computing. (41), pp. 670–678. https://doi.org/10.1007/BF02349975

Oktay M.O. & Akdeniz R. (2014 April). Total Larenjektomi Hastaları için Konusma Kodlama Sistemi. In CD: SIU2014 IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı; Karadeniz Teknik Üniversitesi, Trabzon, Turkey. (in Turkish)

Tarakçıoğlu G.S. (2010). Voice conversion for reconstruction of dysphonic speech. MSc Thesis, Boğaziçi Üniversitesi, Istanbul, Turkey.

Türkmen H.I. (2008). Karma uyarım doğrusal Öngörüm kodlaması yöntemi ile disfonik konuşmadan normal konuşma elde edilmesi. MSc. Thesis, Yıldız Teknik Üniversitesi, Istanbul, Turkey. (in Turkish)

Virag N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. IEEE Transactions on Speech and Audio Processing. 7 (2), pp. 126-137. https://doi.org/10.1109/89.748118