Interspeech 2018
2-6 September 2018, Hyderabad

Reconstructing neutral speech from tracheoesophageal speech

Abinay Reddy N, Achuth Rao M V, G. Nisha Meenakshi, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore- 560012, India

nainireddy@iisc.ac.in, achuthr@iisc.ac.in, nishag@iisc.ac.in, prasantg@iisc.ac.in

Abstract

In this work, we propose a tracheoesophageal (TE) speech to neutral speech conversion system using data collected from a laryngectomee. In laryngectomees, in the absence of vocal folds, it is the vibration of the esophagus that gives rise to a low frequency pitch during speech production. This pitch is manifested as impulse-like noise in the recorded speech. We propose a method to first 'whisperize' the TE speech prior to the linear predictive coding (LPC) based synthesis which uses pitch derived from the energy contour. In order to perform 'whisperization', we model the LPC residual signal as the sum of white noise and impulses introduced by the esophageal vibrations. We model these impulses and white noise using Bernoulli-Gaussian distribution and Gaussian distribution, respectively. The strength and location of the impulses are estimated using Gibbs sampling in order to remove the impulse-like noise from speech to obtain whispered speech. Subjective evaluation via listening test reveals that the 'whisperization' step in the proposed method aids in synthesizing a more natural sounding neutral speech. A different listening test shows that the listeners prefer the synthesized speech from the proposed method $\sim 93\%$ (absolute) times more than the best baseline scheme.

Index Terms: Voice Prosthesis, Laryngectomy, Whispered speech, Tracheoesophageal

1. Introduction

During advanced stages of cancers of the larynx and hypopharynx, a surgical procedure called laryngectomy is performed, in which cancerous regions, including the larynx or the voice box, are removed [1]. Voice rehabilitation is done post laryngectomy in order to help patients produce intelligible speech even in the absence of vocal fold vibrations. In such scenarios, procedures involving the tracheoeseopahgeal puncture (TEP) are preferred among various options for speech production including esophageal speech and artificial electro-larynx [2, 3]. In this procedure, the TEP holds in place a voice prosthesis that functions as a one-way valve to allow the air from the lungs to enter the esophagus. A tracheal stoma (opening in the front of the neck) is made for the laryngectomees to breathe. Therefore, during speech production, the stoma must be closed (typically done using the laryngectomee's thumb or hands-free devices [4]). The air from lungs is allowed by the prosthesis into the esophagus, which vibrates. This air is then shaped by vocal tract and speech is rendered. Thus, the 'voicing' in the tracheoesophageal (TE) speech [5] is due to the vibrations in the cervical esophagus [6]. These vibrations yield an artificial pitch that is characterized by low average fundamental frequency for both male and female laryngectomees, alike [7, 8].

The tracheoesophageal (TE) speech is known to be hoarse, breathy and rough [9, 10]. Although preferred to esophageal and electro-larynx speech [8], TE speech is still considered to be poor in terms of perceptual naturalness compared to neutral (laryngeal) speech [11]. During speech production by laryngectomees, noise due to improper closure of the stoma, called stomal noise, could be introduced [6], further degrading the voice quality. Improper stomal occlusion can cause inability to produce voice. In addition to these, female laryngectomees exhibit a higher degree of voice handicap compared to male patients due to the low fundamental frequency of the TE speech [9]. Therefore, there is a need to convert this TE speech into a more natural sounding neutral speech.

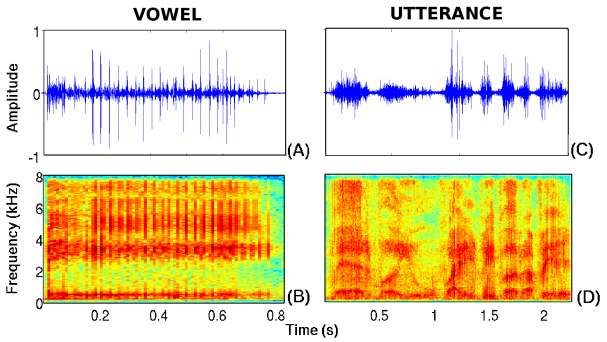

Figure 1: *TE speech signal and corresponding spectrogram for vowel /e:/ (A) and (B) and for an utterance 'Where were you while we were away?' (C) and (D). Spectrograms are computed with a window length of* 20*ms and an overlap of* 10*ms.*

In order to perform such a conversion, we collect speech data from a female laryngectomee. Fig. 1 illustrates speech samples and the corresponding spectrogram for a vowel and for an utterance. From Fig. 1 we observe the manifestation of the esophageal pitch as impulses in the time domain signal ((A) and (C)) with an impulsive broadband structure in the spectro-temporal domain ((B) and (D)). Existing works to convert speech from laryngectomees into neutral speech typically assume that the speech produced is unvoiced or whispered [12, 13, 14]. These methods convert whispered speech into neutral speech without considering the effect of the impulses typically present in the TE speech. In this work, we develop a TE speech to neutral speech conversion system, referred to as TE2N scheme, which first removes the effect of the impulses introduced by the vibration of the esophagus. This step, named the 'whisperization' step, aims to estimate and eliminate the impulses from the TE speech to obtain an unvoiced whispered-like speech. In the second step of the proposed framework, we estimate the pitch from the short-time energy contour of the whisperized speech and synthesize neutral speech using the linear predictive coding (LPC) synthesis. Subjective evaluations reveal the importance of the 'whisperization' step and shows that the proposed method synthesizes a more natural sounding speech compared to the state-of-the-art technique.

10.21437/Interspeech.2018-1907

2. Proposed TE2N scheme

The steps of the conversion system are shown in Fig. 2. The TE2N scheme consists of two main steps: 1) 'Whisperization' : removal of the impulse-like noise introduced by the esophageal vibrations to obtain whisperized speech 2) 'Neutral speech Synthesis' : conversion of this whisperized speech to neutral speech. Each of these steps are described in detail below.

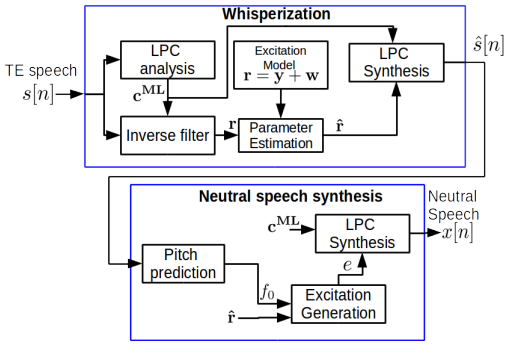

Figure 2: *Proposed tracheoesophageal speech to neutral speech conversion system.*

2.1. Whisperization

2.1.1. Model for tracheospeech generative process

We assume that the observed TE speech samples $s[n]$ is generated by source filter model [15], in which the TE speech signal $s[n]$ is produced by passing the excitation signal through an all-pole filter. We compute the LP coefficients($\mathbf{c^{ML}}$) from the observed speech signal $s[n]$ in a frame of length N_w using the covariance method [16]. The residual of the LPC is given by, $r[n] = s[n] - \sum_{k=1}^{K} c^{ML}[k]s[n-k]$, where K is the order of the all-pole filter.

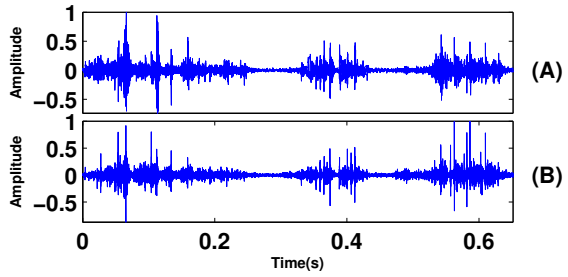

Figure 3: *Comparison of (A) TE speech and (B) corresponding LPC residual.*

Fig. 3 shows the TE speech and corresponding LP residual. It can be observed from the figure that the LP residual is characterized by impulse like temporal events. Hence we assume that the LP residual a sum of impulses (y) and white Gaussian noise (w), where the impulse like noise is introduced by the esophageal vibrations. Thus the residual, $\mathbf{r}$, is modeled as,

$$\mathbf{r} = \mathbf{y} + \mathbf{w} \tag{1}$$

where $\mathbf{r} = [r[0], r[1], \ldots, r[N_w - 1]]^T$, $\mathbf{y} = [y[0], y[1], \ldots, y[N_w - 1]]^T$, $\mathbf{w} = [w[0], w[1], \ldots, w[N_w - 1]]^T$, $\mathbf{w} \sim \mathcal{N}(w; 0, \sigma^2 I)$ and $\mathcal{N}(\mu, \Sigma)$ indicates the Gaussian distribution with mean μ and covariance Σ. Thus, to eliminate the impulses, we first estimate the impulses, $\mathbf{y}$, given the LP residual $\mathbf{r}$.

2.1.2. LPC residual Model

We model the impulses in the residual signal $\mathbf{y}$ given in Eq. 1 using a product of a binary random variable $\mathbf{b}$ and a Gaussian random variable $\mathbf{a}$ ($y[k] = a_k \times b_k$). Eq. 1 can be rewritten as $\mathbf{r} = A\mathbf{b} + \mathbf{w}$. Let $\mathbf{a} = [a_0, a_1, \ldots, a_{N_w-1}]^T$, $\mathbf{b} = [b_0, b_1, \ldots, b_{N_w-1}]^T$ where A is a diagonal matrix with the diagonal entries as in $\mathbf{a}$. Thus, the set of parameters of the model is $\mathbf{\Theta} = \{\mathbf{a}, \mathbf{b}, \sigma^2\}$. $\mathbf{\Theta}$ is estimated given the residual signal $\mathbf{r}$. The likelihood of the given residual signal $\mathbf{r}$ is given by

$$p(\mathbf{r}|\mathbf{\Theta}) = \frac{1}{\sqrt{(2\pi(\sigma^2)^{N_w})}} e^{-\frac{(\mathbf{r}-A\mathbf{b})^T(\mathbf{r}-A\mathbf{b})}{2\sigma^2}} \tag{2}$$

In general, the number of parameters $(2N_w + 1)$ are more than that (N_w) of the observed residual samples $\mathbf{r}$ for parameter estimation. Hence, we impose a prior distribution on each of the model parameters. The details of the priors are described next.

2.1.3. Parameter priors

a. Residual prior: In the model of the residual signal (Eq. 1), we assume that the a_k and b_k are independent of $a_{l\neq k}$ and $b_{l\neq k}$. The distribution of y is given by

$$p(\mathbf{a}, \mathbf{b}) = \prod_{k=0}^{N_w-1} p(a_k, b_k) = \prod_{k=0}^{N_w-1} p(a_k|b_k)p(b_k)$$

where

$$p(b_k) = Be(b_k; \lambda), \tag{3}$$

$Be(z; \lambda)$ indicates the Bernoulli distribution with parameter λ [17]. $a_k|b_k$, $\forall k$ is assumed to have identical distribution as follows,

$$p(a_k|b_k) = \begin{cases} \mathcal{N}(a_k; 0, \sigma_a^2) & \text{, if } b_k = 1 \\ \delta(a_k) & \text{, if } b_k = 0 \end{cases} \tag{4}$$

where σ_a^2 is the variance and $\delta(\cdot)$ denotes the Dirac delta function.

b. Noise variance prior: The prior distribution of noise variance is assumed to be inverse Gamma distribution with parameters α and β.

$$p(\sigma^2) = \mathcal{IG}(\sigma^2; \alpha, \beta) \tag{5}$$

where $\mathcal{IG}(z; \alpha, \beta)$ indicates the inverse Gamma distribution with shape parameter α and scale parameter β [18]. We assume that the random variables (a_k, b_k) representing the excitation signal, and noise variances (σ^2) are independent of each other. The joint prior of the parameters is given by

$$p(\mathbf{\Theta}) = p(\sigma^2) \prod_{k=0}^{N_w-1} p(a_k|b_k)p(b_k) \tag{6}$$

2.1.4. Parameter estimation

Given the likelihood of residual signal in Eq. 2 and the model parameters prior $p(\mathbf{\Theta})$ in Eq. 6, the posterior distribution of the parameters is given by

$$p(\mathbf{\Theta}|\mathbf{r}) \propto p(\mathbf{r}|\mathbf{\Theta})p(\mathbf{\Theta}) = p(\mathbf{r}|\mathbf{\Theta}) \prod_{k=0}^{N_w-1} p(a_k|b_k)p(b_k)p(\sigma^2). \tag{7}$$

The parameters $\mathbf{\Theta}$ are estimated by maximizing the posterior distribution in Eq. 7 as follows,

$$\{\mathbf{a}^*, \mathbf{b}^*, \sigma^{2*}\} = \arg\max_{\mathbf{a},\mathbf{b},\sigma^2} \quad p(\mathbf{\Theta}|\mathbf{r}). \tag{8}$$

But directly maximizing the posterior distribution in Eq. 7 is, in general, intractable. So we use Gibbs sampling to generate the samples from the posterior distribution and then estimate the parameters. The Gibbs sampling [19] is a method, which is used to generate the samples from the joint distribution by iteratively sampling from the conditional distributions. Hence we derive the conditional distribution with respect to each model parameter given other parameters and the speech samples. The joint distribution of a_k, b_k can be written as a product of two conditional distributions, $p(\{a_k, b_k\}|\mathbf{\Theta} \setminus \{a_k, b_k\}, \mathbf{r}) = p(b_k|\mathbf{\Theta} \setminus \{a_k, b_k\}, \mathbf{r})p(a_k|\mathbf{\Theta} \setminus \{a_k\}, \mathbf{r})$, where $\mathbf{\Theta} \setminus \{\tau\}$ indicates all parameters $\mathbf{\Theta}$ except the parameter τ. To get the sample from $p(\{a_k, b_k\}|\mathbf{\Theta} \setminus \{a_k, b_k\}, \mathbf{r})$, we first sample from $p(b_k|\mathbf{\Theta} \setminus \{a_k, b_k\}, \mathbf{r})$ and then sample from $p(a_k|\mathbf{\Theta} \setminus \{a_k\}, \mathbf{r})$. The required conditional distribution for each model parameter is as follows (the derivations are shown in the supplementary material)

$$p(b_k|\mathbf{\Theta} \setminus \{a_k, b_k\}, \mathbf{r}) = Be\left(b_k; \frac{\lambda_{1,k}}{\lambda_{1,k} + 1 - \lambda}\right) \tag{9}$$

$$p(a_k|\mathbf{\Theta} \setminus \{a_k\}, \mathbf{r}) = \mathcal{N}\left(\frac{d_m}{f}, \frac{1}{f}\right) \tag{10}$$

$$p(\sigma^2|\mathbf{\Theta} \setminus \{\sigma^2\}, \mathbf{r}) = \mathcal{IG}\left(\alpha + \frac{N_w}{2}, \frac{1}{2}\mathbf{d_3}^T\mathbf{d_3} + \beta\right) \tag{11}$$

where $\lambda_{1,m} = \frac{\lambda}{\sigma_a}\sqrt{\frac{1}{f}}e^{\left(\frac{d_m^2}{2f}\right)}$, $\mathbf{d_3} = \mathbf{r} - A\mathbf{b}$ and $f = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_a^2}\right)$. We sample from the distributions in Eq. 9, Eq. 10 and Eq. 11. N_i number of times to get the samples $\{b_k[i], a_k[i], \sigma^2[i], 0 \leq i \leq N_i, 0 \leq k \leq N_w\}$. We estimate the optimum value of a_k and b_k as follows,

$$a_k^* = \frac{1}{N_i - N_b}\sum_{i=N_b}^{N_i} a_k[i], \quad b_k^* = \arg\max_{b_k \in \{0,1\}} q(b_k)$$

where N_b is the number of burn-in iterations [19] and $q(b_k)$ is the relative multiplicity of the $b_k \in \{0, 1\}$ in the samples $b_k[i]$. Using a_k^* and b_k^*, $y^*[k] = a_k^* \times b_k^*$ is computed. The estimate of the impluse-like noise, $y^*[k]$ is removed from the residual signal $\mathbf{r}$ to get the whisperized speech residual $\hat{r}[k] = r[k] - y^*[k]$. The whisperized speech residual passed through the LP filter $\mathbf{c}^{ML}$ to get the whisperized speech.

$$\hat{s}[k] = \hat{r}[k] + \sum_{j=1}^{K} \hat{s}[k-j]c^{ML}[j] \tag{12}$$

2.2. Neutral Speech synthesis

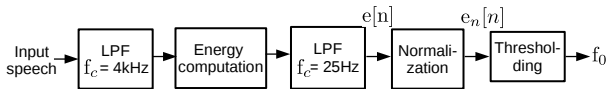

Figure 4: *Block diagram describing pitch prediction. f_c denotes cut-off freqeuncy of the low-pass filter (LPF)*

Fig. 2 provides the steps of the neutral speech synthesis which are described below: Pitch Prediction: Fig. 4 provides the steps involved in predicting pitch. First we compute the short-term energy contour of the low-pass filtered (cut off frequency 4kHz) whisperized speech with frame length of 25ms and shift of 10ms and then smoothen the energy contour to remove large variations to get $e[n]$. In order to achieve a standard deviation (SD) of 10 (empirically chosen) for the predicted pitch contour, we normalize the energy contour to obtain, $e_n[n] = 10 \times e[n]/\sigma_e$, where σ_e is the SD of $e[n]$. To compute the final pitch contour f_0, we perform a thresholding as follows,

$$f_0[n] = \begin{cases} e_n[n] + \bar{f}_0 & \text{, if } e[n] > \epsilon \\ 0 & \text{, otherwise} \end{cases} \tag{13}$$

$\bar{f}_0$ indicates the subject-specific average pitch. We set $\bar{f}_0$ to a typical average female pitch of 220Hz. f_0=0 corresponds to predicted unvoiced segment. We find that using such a scheme of thresholding helps in avoiding a separate voice/unvoiced detector.

Excitation generation: Given the f_0, we generate the impulses similar to [20]. We add a high pass filtered ($f_{cutoff} = 1$kHz) residual ($\hat{\mathbf{r}}$) to the impulses to get the final excitation signal (e). In addition to reducing the buzz in the speech [20], this also acts as the excitation signal in the unvoiced regions.

LPC synthesis: The excitation signal (e) is filtered through the LP filter ($\mathbf{c}^{ML}$) to get the neutral speech $x[n]$ in each frame. The final speech is generated from frames using overlap add method.

3. Dataset

We recorded data from a 70 year old female laryngectomee who uses the AUM voice prosthesis [21]. The subject is proficient in reading, writing and speaking in English. Although analysis of pathological speech is typically done using sustained vowel phonation, the use of connected speech is known to be useful since it reflects the usual context of everyday [10]. Therefore, we collected data that includes sustained vowel phonation, voiced utterances and questions. Specifically, we recorded five repetitions of the vowels, /a:/, /i:/, /u:/, /e:/ and /o:/. A list of 29 voiced sentences was also designed (the complete list is provided in the supplementary material[1]). In addition to these, a set of 21 phonetically balanced utterances, from the MOCHA-TIMIT database [22], that corresponded to questions were also recorded. The recordings were carried out using the Zoom H6 recorder at a sampling rate of 44100Hz and were later downsampled to 16kHz for further analysis.

4. Experiments

For the experiments we consider the voiced sentences and the questions resulting in a total of 50 utterances. Speech is divided into overlapping frames of length 25ms ($N_w = 400$) and shift of 10ms. We use $K = 15$, $\alpha = \frac{N_w}{4}$, $\beta = 0$, $\lambda = 0.03$, $\sigma_a = 1$ and $\epsilon = 10^{-6}$. We found that $N_i = 40$ is sufficient for convergence and $N_b = 10$ is sufficient to get samples from the posterior distribution. We perform two experiments– 1) to understand the significance of the 'whisperization' step and compare the performance with two baseline schemes, namely, B1 and B2, that do not whisperize the TE speech prior to synthesis; 2) to evaluate the performance of the proposed TE2N method in comparison to the existing baseline schemes namely, B3 and B4. We now describe the four baseline schemes.

4.1. Baseline Schemes

In first baseline scheme (B1), we reconstruct neutral speech from TE speech without whisperizing it prior to the neutral speech synthesis (Fig. 2). In order to eliminate the effect of any impulse-like excitation in the original TE speech, we consider a second baseline scheme B2, similar to B1, that substitutes the residual $\hat{r}$, with white Gaussian noise ($\mathcal{N}(0,1)$). In both B1 and B2 the pitch is derived from the energy contour of $s[n]$ (Fig. 4). The third and fourth baseline schemes, B3 and B4, are motivated by the state-of-the-art whisper to neutral speech conversion scheme proposed by Mcloughlin *et. al,* [14]. While B3 takes the TE speech, directly as input, B4 considers the whisperized speech as input.

4.2. Subjective Evaluation

Since the recordings of the original voice of the subject prior to laryngectomy are unavailable, direct evaluation of the performance of the different schemes using objective measures becomes challenging. Hence, subjective evaluations via two listening tests are performed. In the first listening test (LT-I), we examine the importance of the 'whisperization' step by comparing the performance of the proposed scheme and the two baseline schemes, B1 and B2. We synthesize 50 utterances using the three schemes (B1, B2, TE2N) and presented them to each listener. The listeners were asked to choose the best synthesis scheme based on naturalness. The listeners could choose more than one of the three schemes if they sounded equally natural. Since both the baseline schemes do not whisperize the original TE speech, the scheme with the highest preference would highlight if the 'whisperization' step is, indeed, beneficial. In the second listening test (LT-II), we compare the performance of the proposed method with the two baseline schemes B3 and B4 that are based on the state-of-the-art reconstruction technique for speech from laryngectomees [14]. In case, all three schemes (B3, B4, TE2N) resulted in a poor sounding synthesized speech, the listeners could choose the 'none' option. In order to aid the listeners, the text corresponding to the utterance being presented was also provided in the graphical user interface (designed using MATLAB R2014a) in both the listening tests.

In each listening test, we ensured that each utterance is evaluated by 5 listeners. Hence, we chose 10 listeners (5 males and 5 females) who are proficient in reading, writing and speaking English to perform both the subjective evaluations. The average age of the listeners was $23.70(\pm 1.06)$ years. It was ensured that the utterances presented to a listener during one test were not repeated in the other test. In order to check if the listeners were consistent in their choices, ten utterances were repeated during the course of the evaluation. All listeners turned out to be at least 70% and 80% consistent in LT-I and LT-II, respectively. The average time taken to complete LT-I and LT-II was $19.05(\pm 5.38)$ minutes and $9.57(\pm 3.08)$ minutes, respectively.

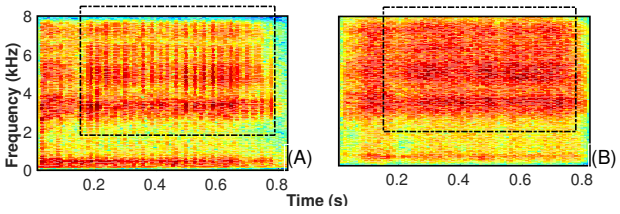

Figure 5: *Spectrograms of (A) TE speech and (B) whisperized speech for the vowel /e:/.*

Table 1: *Preference Scores from LT-I*

Methods	TE2N	B1	B2	B1,B2
Preference (in %)	**62.4**	16.4	15.6	3.6
Methods	TE2N,B1	TE2N,B2	TE2N,B1,B2	
Preference (in %)	1.2	0	0.8	

Table 2: *Preference Scores from LT-II*

Methods	TE2N	B3	B4	None
Preference (in %)	**94.40**	0.8	0	4.80

5. Results and Discussion

Fig. 5 shows the effect of the 'whisperization' step for vowel /e:/. From the figure we observe that the impulsive broadband structure due to the esophageal vibrations are eliminated in the whisperized speech (black dashed boxes in the figure). This highlights the effectiveness of the proposed 'whisperization' step to obtain an impulse noise free unvoiced speech from TE speech. Results from LT-I reveal that this step is vital to synthesize natural sounding neutral speech. Table 1 provides the preference scores corresponding to LT-I. From the table, it is evident that the proposed TE2N scheme, is preferred as the best 46% (absolute) times more than the best baseline scheme, B1. Among the baseline schemes, we observe that B2 is preferred to B1 for questions. It could be that the absence of impulse-like excitation in the unvoiced regions in B2, yields a better sounding neutral speech compared to B1. Interestingly, we find that while the overall performance of the two baseline schemes are comparable, they differ from that of the proposed method. This reveals that significant improvement in naturalness is obtained after whisperizing the original TE speech. The preference scores corresponding to LT-II are provided in Table 2 [1]. From the table we find that the proposed method outperforms the baseline schemes. The poor performance of the baseline schemes could be due to the modification of the spectrum and computation of pitch using formants, the estimation of which could be affected by the nature of TE speech. These results confirm that the proposed TE speech to neutral speech conversion system synthesizes a more natural sounding neutral speech compared to the existing techniques.

6. Conclusion

In this work, we propose a tracheoesophageal (TE) speech to neutral speech conversion system that first whisperizes the TE speech by eliminating the impulse-like noise and then performs LPC based synthesis. Using listening tests, we find that the 'whisperization' step is vital to improve the naturalness of the synthesized speech compared to several baseline schemes. Our future work includes data collection from many laryngectomees and extending the proposed approach to the same.

7. Acknowledgements

We thank the subject, Dr. Manjula B.V and Sahana Monhandoss for the data collection and Pratiksha Trust for their support.

[1] Derivations, list of sentences and examples of speech reconstructed using different schemes are available at https://spire.ee.iisc.ac.in/spire/software.php

8. References

[1] J. Fagan, "Open access atlas of otolaryngology, head & neck operative surgery," University of Cape Town. [Online]. Available: https://vula.uct.ac.za/access/content/user/01372298/Total %20laryngectomy.pdf

[2] M. K. El-Sharnobya, E. A. Behairya, A. A. Abdel-Fattah, M. A. Al-Belkasy *et al.*, "Voice rehabilitation after total laryngectomy," *Menoufia Medical Journal*, vol. 28, no. 4, pp. 800–806, 2015.

[3] M. I. Singer and E. D. Blom, "Tracheoesophageal puncture: A surgical prosthetic method for postlaryngectomy speech restoration," in *Third International Symposium on Plastic-Reconstructive Surgery of the Head and Neck, New Orleans*, vol. 4, 1979.

[4] E. Houwen, "Development of a handsfree speech valve for laryngectomy patients," Ph.D. dissertation, 2012, relation: https://www.rug.nl/ Rights: University of Groningen.

[5] M. I. Singer, "Tracheoesophageal speech: vocal rehabilitation after total laryngectomy," *Laryngoscope*, vol. 11, no. 1, pp. 1454–1465, 1993.

[6] H. F. Nijdam, A. A. Annyas, H. K. Schutte, and H. Leever, "A new prosthesis for voice rehabilitation after laryngectomy," *Archives of oto-rhino-laryngology*, vol. 237, no. 1, pp. 27–33, Dec 1982. [Online]. Available: https://doi.org/10.1007/BF00453713

[7] Y. Qi, B. Weinberg, and N. Bi, "Enhancement of female esophageal and tracheoesophageal speech," *The Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2461–2465, 1995.

[8] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, "A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production," *Journal of Speech and Hearing Disorders*, vol. 49, no. 2, pp. 202–210, 1984. [Online]. Available: http://dx.doi.org/10.1044/jshd.4902.202

[9] R. Kazi, E. Kiverniti, V. Prasad, R. Venkitaraman, C. Nutting, P. Clarke, P. RhysEvans, and K. Harrington, "Multidimensional assessment of female tracheoesophageal prosthetic speech," *Clinical Otolaryngology*, vol. 31, no. 6, pp. 511–517, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2273.2006.01290.x

[10] A. Singh, R. Kazi, J. D. Cordova, C. Nutting, P. Clarke, K. Harrington, and P. RhysEvans, "Multidimensional assessment of voice after vertical partial laryngectomy: A comparison with normal and total laryngectomy voice," *Journal of Voice*, vol. 22, no. 6, pp. 740 – 745, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0892199707000616

[11] M. H. Bellandese, J. W. Lerman, and H. R. Gilbert, "An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 6, pp. 1315–1320, 2001. [Online]. Available: + http://dx.doi.org/10.1044/1092-4388(2001/102)

[12] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[13] J. j. Li, I. V. McLoughlin, L. R. Dai, and Z. h. Ling, "Whisper-to-speech conversion using restricted Boltzmann machine arrays," *Electronics Letters*, vol. 50, no. 24, pp. 1781–1782, 2014.

[14] I. V. Mcloughlin, H. R. Sharifzadeh, S. L. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Trans. Access. Comput.*, vol. 6, no. 4, pp. 12:1–12:21, May 2015. [Online]. Available: http://doi.acm.org/10.1145/2737724

[15] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, 1971, vol. 2.

[16] L. R. Rabiner, R. W. Schafer *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[17] J. V. Uspensky, "Introduction to mathematical probability," 1937.

[18] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," Tech. Rep., 2007.

[19] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 721–741, 1984.

[20] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[21] I. M. Devices. AUM voice prosthesis. [Online]. Available: http://www.innaumation.com/

[22] A. Wrench, "MOCHA-TIMIT," Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database, 1999. [Online]. Available: http://sls.qmuc.ac.uk