Interspeech 2018
2-6 September 2018, Hyderabad

Designing a Pneumatic Bionic Voice Prosthesis: A Statistical Approach for Source Excitation Generation

Farzaneh Ahmadi[1], *Tomoki Toda*[2]

[1]MARCS Institute, Western Sydney University
[2]Information Technology Center, Nagoya University
f.ahmadi@uws.edu.au, tomoki@icts.nagoya-u.ac.jp

Abstract

This study follows up on our pioneering work in designing a Pneumatic Bionic Voice (PBV) prosthesis for larynx amputees. PBV prostheses are electronic adaptations of the traditional Pneumatic Artificial Larynx (PAL) device. The PAL is a non-invasive mechanical voice source, driven exclusively by respiration and with an exceptionally high voice quality. Following the PAL design closely, the PBV prosthesis is anticipated to substitute the medical gold standard of voice prostheses by generating a similar voice quality while remaining non-invasive and non-surgical. This paper describes a statistical approach to estimate the excitation waveform of the PBV source using the PAL as a reference. A Gaussian mixture model of the joint probability density of respiration and PAL voice features is implemented to estimate the excitation waveform of the PBV. The evaluation on a database of more than two hours of continuous speech shows a close match between f_0 pattern and mel-cepstra of the estimated PBV source and the PAL. When used to re-synthesize the original speech, the intelligibility of the PBV speech remains high and is scored 7.1±0.4 compared to 7.9±0.15 of the original PAL source.

Index Terms: Bionic Voice, artificial larynx, laryngectomy, pneumatic larynx, voice synthesis

1. Introduction

Total laryngectomy is the surgical removal of the larynx due to cancer which leads to a lifetime of voice-loss. Despite the emergent progress in many fields of bionics, a functional Bionic Voice solution still does not exist for laryngectomy patients. In a source-filter model of speech generation, a Bionic Voice prosthesis substitutes only the voice source and provides the voice generation function of the missing vocal folds for the patient. The patient next shapes this artificial source excitation into speech by moving their face and lips muscles.

Among available solutions to generate voice after the laryngectomy, the medical gold standard of Tracheoesophageal (TE) voice prosthesis [1] and the Pneumatic Artificial Larynx (PAL) devices [2] continue to generate a superior voice quality and outperform other electronic voice prostheses including the Electrolarynx [3] and silent speech interfaces [4] both in terms of intelligibility and naturalness [6–13]. The gold standard of TE voice prosthesis is a plastic valve placed surgically inside the throat which subjects the patient to infection bio-hazards [1]. The PAL is specifically of interest as a non-invasive respiratory driven, mechanical voice source with an exceptionally high voice quality [2, 5-8]. When a laryngectomy patient loses their larynx, an opening in the neck (stoma) is generated for them to breathe. The PAL is placed externally between this stoma and the mouth and is driven by the variations of the pressure at these two points. In that sense, the PAL is essentially a mechanical model of human larynx with a fixed pair of vocal folds, driven exclusively by respiration, which generates the excitation source signal of the speech. Figure 1 shows the signal flow from respiration to PAL source excitation when a patient uses the PAL as a voice prosthesis.

The PAL holds strong advantages against the gold standard of TE prosthesis by being non-invasive, having a clearer voice, less noise and higher intelligibility levels [2, 5,7-11]. Yet, the PAL has limited prevalence due to its cumbersome design. The Pneumatic Bionic Voice (PBV) is proposed as an electronic adaptation of the traditional PAL to eliminate its shortcoming while using voice conversion to mimic its voice generation function [12]. Hence, the PBV is expected to substitute the existing gold standard by producing a voice quality similar to the PAL while remaining non-invasive and non-surgical.

To design a PBV source, mimicking the PAL's performance, we need to answer two questions:1) how to control the onset and offset of the PBV source similar to the PAL, 2) how to estimate the voice generation function of the PAL driven by respiration and generate a similarly high voice quality for the PBV source? We tried answering the first question in our previous work [12]. This study aims to answer the second and implements a statistical framework to estimate the PAL source excitation generation from respiration. A respiration to voice conversion system is designed based on two major statistical conversion approaches [13, 14]. The system is trained and tested on a dataset of voice and respiration, recorded from a laryngectomy patient using the PAL to speak. Next, objective and subjective evaluation of the method is performed through re-synthesizing the speech signal, substituting the estimated PBV source waveform with the original PAL.

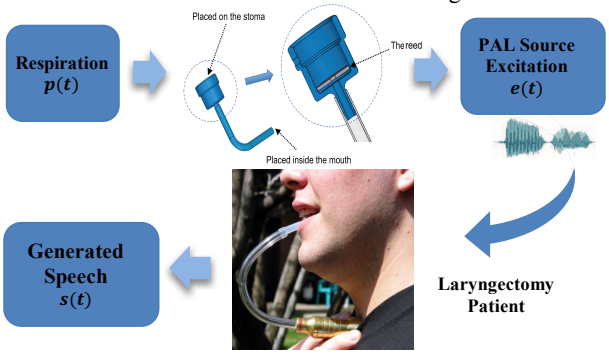

Figure 1: *Signal flow of a laryngectomy patient using the PAL to speak. The patient's respiration p(t) drives the vibrations of the PAL reed and generates the source excitation signal e(t). When transferred to the oral cavity via a tube, e(t) excites the vocal tract and generates the speech signal s(t).*

10.21437/Interspeech.2018-1043

2. Statistical approach for estimation of PAL source excitation

To Estimate the PAL source excitation from respiration signal, the general statistical voice conversion (VC) framework [14] is employed. In a traditional VC system, the speech waveform of one speaker is converted to the speech of the other. Hence, the feature sets of choice at the two sides of the conversion are the same. Here, a slow varying respiration signal at one side of the conversion, should be converted to the feature sets of a PAL's excitation voice source at the other. So, in addition to the choice of the statistical conversion framework, careful consideration has to be made in deciding which features should be extracted from the respiration and PAL voice source signals.

2.1. A VC framework to convert respiration to PAL source excitation

This study employs the framework of using Gaussian Mixture Models to convert a trajectory of features of the source (respiration signal) to the target (PAL excitation waveform). Considering the PAL as a black box, with respiration signal $p(t)$ as the input and the PAL excitation waveform $e(t)$ as the output, the underlying mapping function $\hat{e} = f(p)$ is estimated as follows. A trajectory (time sequence) of the features of the respiration (p) and PAL voice (e) are shaped. Given P_t and E_t as the D_x, D_y-dimensional respiration and PAL voice feature vectors at frame *t*, respectively, the trajectories are defined as: $\boldsymbol{P} = [{P_1}^T\ {P_2}^T, \dots {P_t}^T, \dots {P_\mathcal{T}}^T]^T$, $\boldsymbol{E} = [{E_1}^T\ {E_2}^T, \dots {E_t}^T, \dots {E_\mathcal{T}}^T]^T$. The notation $.^T$ denotes transposition of a vector.

A Gaussian Mixture Model is trained to describe the joint probability distribution of these trajectories. Next, the joint probability density of the source and target feature vectors $\boldsymbol{Z_t} = [\boldsymbol{P}^T, \boldsymbol{E}^T]^T$ is described by a GMM as:

$$\mathbb{P}(\boldsymbol{Z_t}|\lambda) = \sum_{m=1}^{\boldsymbol{M}} \omega_m \mathcal{N}\,(\boldsymbol{Z_t}; \mu_m^{(z)}, \Sigma_m^{(z)}) \quad (1)$$

with $\mathcal{N}(.;\mu,\Sigma)$ being a normal distribution (with μ and Σ) and λ being the parameter set of the GMM which consists of weights, mean vectors and covariance matrices of each mixture component (with index m) and a total of $\boldsymbol{M}$ mixture components. The parameters of the GMM are fitted to the training dataset of respiration and voice feature trajectories using EM algorithm [15].

Once trained, to perform the conversion, given any new respiration feature vector $\boldsymbol{P}$, a maximum likelihood approach determines the corresponding $\widehat{\boldsymbol{E}}$ using on the estimated joint probability function.

$$\widehat{\boldsymbol{E}} = \text{argmax}\ \mathbb{P}(\boldsymbol{E}|\boldsymbol{P}, \lambda) \quad (2)$$

In the conversion process, the converted voice source feature $\widehat{\boldsymbol{E}}$ is estimated from the respiration features trajectory $\boldsymbol{P}$ in the same manner as maximum likelihood estimation of parameter trajectories with the GMM [14].

2.2. Feature selection from respiration and voice signals

The PAL excitation source can be described using three components of spectral envelope, f_0 and aperiodicity, expected with the World vocoder [16]. The spectral envelope, can be further parameterized into the 1-24th mel-cepstral coefficients. These features are not however applicable to extract information from the respirations signal. Respiratory signals that drive the PAL are slow varying with a bandwidth of 0-50 *Hz*, sampled at 1 *kHz*.

To propose a suitable feature set for these, the most informative feature sets of the slow varying myoelectric muscle activity signals [17-19] are consulted. These features have already proven feasible in estimating the fundamental frequency of a Bionic Voice source using myoelectric signals [20]. They had also proven effective in onset and offset detection of the PBV source from respiration.

The PAL voice generation is known to be influenced by pressure variations at its two ends, namely the pressure inside the mouth $p_m(t)$, at the stoma $p_s(t)$ and also majorly by the difference of the two ($\Delta p(t) = p_s(t) - p_m(t)$) [12]. Hence, initially, a subset of effective myoelectric features [19] was extracted from these three signals ($\Delta p(t), p_s(t), p_m(t)$). Next, a feature selection was applied to rule out weak features from this subset, based on their performance in estimating the fundamental frequency (f_0) of the PAL source. To perform the feature selection, a linear regression model was fitted to different subsets of these features to estimate f_0. The criteria for performance evaluation was the correlation coefficient ($R(\hat{f}_0, f_0)$) with $\hat{f}_0$ and f_0 being the estimated and intended f_0. The subsets of features with minimum $R(\hat{f}_0, f_0)$ were eliminated. This criterion revealed the following time and frequency domain features as features of choice for the respiration signal in Eqs. (3)-(9). In these series, Eqs. (3)-(5) were used to extract features from individual stoma and mouth signals ($p_m(t)$ and $p(t)$). Yet all of the Eqs. (3)-(9) were applied to extract features from the pressure difference signal $\Delta p(t)$.

Let $p(t)$ to represent any of the three respiration signals ($\Delta p(t)$, $p_s(t)$, $p_m(t)$) which affect the PAL voice generation. To estimate time domain features, $p(t)$ is segmented into overlapping frames of length *N* samples. For each frame p_t, the time domain features are: the mean absolute value (MAV):

$$MAV = \frac{1}{N}\sum_{i=1}^{N}|p_i| \quad (3)$$

the root mean square (RMS) value:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|p_i|^2} \quad (4)$$

waveform length (WL):

$$WL = \frac{1}{N}\sum_{i=1}^{N}|\Delta p_i| \ \text{ where } \Delta p_i = p_i - p_{i-1} \quad (5)$$

the simple square integral (SSI):

$$SSI = \sum_{i=1}^{N}|p_i|^2 \quad (6)$$

the Modified Mean Absolute Value is also a time domain feature which uses a non-uniform window:

$$MMAV = \frac{1}{N}\sum_{i=1}^{N} w_i|p_i|$$
$$w_i = \begin{cases} 1 & if \quad 0.25N \le i \le 0.75N \\ 0.5 & otherwise \end{cases} \quad (7)$$

and the autocorrelation between adjacent frames

$$\Re(p^J, p^{J+1}) = \frac{1}{N}\sum_{i=1}^{N} p_i^J . p_i^{J+1} \quad (8)$$

where p^J, p^{J+1} are the J^{th} and $(J+1)^{\text{th}}$ frames of $p(t)$ with the length of N.

In the frequency domain the Modified Frequency MedDian (MFMD) (9) passed the feature selection criterion.

$$MFMD = 1/2\ \sum_{j=1}^{M} A_j \quad (9)$$

where A_j is the amplitude of the respiration signal spectrum P(*f*) at frequency bin j and *M* is the total number of frequency bins.

2.3. Implementation details

Figure 2 summarizes the proposed respiration to PAL source excitation conversion system. The PAL's driving pressure signals ($p_m(t)$, $p_s(t)$ and their difference: $\Delta p(t)$) are the input and the PAL's generated voice (PAL source excitation $e(t)$) is the intended output of the conversion system. The core framework of the system has been adopted from the statistical unvoiced speech enhancement method [13]. The PAL intended output $e(t)$, is analysed and synthesized using World [16]. To re-synthesize $e(t)$ using respiration input signals, parameters of aperiodicity the f_0 and mel-cepstrum coefficients of $e(t)$, need to be estimated. In the training phase, two GMMs are trained for f_0 and mel-cepstrum estimation using joint probability density of respiration and voice feature trajectory (1). The trained GMMs are used in the test phase to estimate f_0 and mel-cepstrum coefficients of $e(t)$, using respiration features and to resynthesize the estimated PAL excitation. The aperiodicity parameter is set to constant in this synthesis.

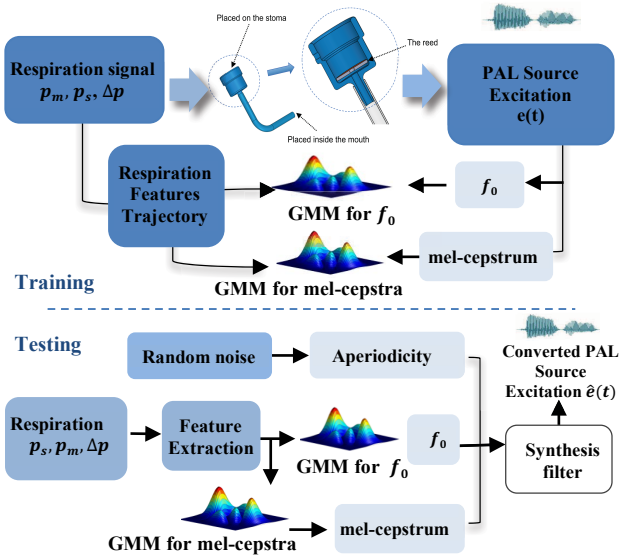

Figure 2: *The GMM based respiration to PAL excitation source conversion.*

3. Methodology of experimental evaluation

3.1. Experimental conditions

The recording condition and the pre-processing of the data was similar to our previous study [12]. A laryngectomy patient who is a proficient user of the PAL device sat in a quiet room and used the PAL as his voice prosthesis to generate continuous speech. The pressures inside his mouth $p_m(t)$ and in front of his stoma $p_s(t)$ were recorded using pressure sensors connected to two thin plastic tubes placed intra-oral and in front of the stoma. These, were the inputs of the respiration to PAL conversion system. A microphone was implanted inside the PAL source to record the PAL excitation source signal $e(t)$ as the intended PAL output. A second microphone recorded the patient' PAL speech $s(t)$ for evaluation purposes. All audio and respiration channels where recorded simultaneously and were time-aligned while recording. The audio ($s(t)$, $e(t)$) and respiration signals ($p_s(t)$, $p_m(t)$) were recorded at 1 *kHz* and 16 *kHz* respectively.

More than two hours of continuous speech and respiration were recorded. The respiration recordings were low-pass filtered to preserve the 0-40 *Hz* and to eliminate the PAL source vibrations (which had a centre frequency of 110±30 *Hz*) [12]. Next, they were segmented into frames of 10 *ms* length, with 5 *ms* of overlap. For each respiration frame the feature sets in section 2.2 were extracted and a trajectory of features for 5 frames before and after the current frame was generated (25 *ms* lead and lag time). The respiration feature space was subjected to the Principal Component Analysis (PCA) to reduce the dimensions of the space [21] while maintaining 99.9% of the variance of the respiration feature space. The target PAL voice was also segmented into windows of 25 *ms* length with a hop size of 5 *ms*. The target mel-cepstra and target f_0 were calculated using World [16] and Reaper [22], respectively.

The GMMs for f_0 and mel-cepstrum estimation were composed of 10 and 14 Gaussian components respectively. The training was performed on a random selection of 80% of the speech utterances [15]. The trained GMMs were then used on the 20% untrained data (different utterances from the training set) to estimate the PAL excitation source.

3.2. Evaluation framework

Three criteria were proposed to evaluate the performance of the system in estimating f_0 and mel-cepstra of the intended PAL source or as a whole. The f_0 estimation was evaluated based on the correlation coefficient $R(\hat{f}_0, f_0)$ of the estimated $\hat{f}_0$ and target f_0, in accord with previous works [23, 24]. The criterion for evaluation of mel-cepstra was the mel-cepstral distortion between the estimated and target:

$$\text{MelCD [dB]} = 10/\ln 10 \sqrt{2 \sum_{i=1}^{24} (mc_i^{(y)} - \hat{mc}_i^{(y)})^2} \qquad (10)$$

with $mc_i^{(y)}$, $\hat{mc}_i^{(y)}$ being the i^{th} coefficient in the target and estimated mel-cepstra respectively. The PAL reed does not vibrate in unvoiced phonemes as the pressure inside the mouth becomes large [25] and the pressure difference between stoma and mouth (Δp) is too small to excite the reed. Hence both f_0 and mel-cepstra evaluations are performed for voiced speech only.

To evaluate the system as a whole, the estimated PAL source should be used to re-synthesize a copy of the speech signal. To do this, ideally the PAL source has to be applied to the vocal tract filter of the patient to generate the speech signal. But how should we estimate the vocal tract filter in this scenario? Our recoded dataset had an interesting advantage which enabled us to effectively approach this. Two microphones recorded both the original PAL voice source $e(t)$ and the resulting speech signal $s(t)$ which were time aligned. In the frequency domain, this provides the vocal tract function by a division of the short-term Fourier spectrums $V(f)$ =S(f)/E(f) where V(f) is the frequency response of the vocal tract and S(f), E(f) are the corresponding frequency spectrum of the speech and PAL source. This division turns into a subtraction in the log-frequency domain of mel-cepstra as:

$$mc_i^{(V)} = mc_i^{(S)} - mc_i^{(E)} \qquad (11)$$

with, $mc_i^{(S)}$ and $mc_i^{(E)}$ being the i^{th} mel-cepstrum coefficient of the time aligned PAL speech and PAL source respectively and $mc_i^{(V)}$ the i^{th} mel-cepstrum coefficient of the vocal tract. Substituting the converted PAL source $\hat{e}(t)$ for original $e(t)$ in (11), the mel-cepstra of the converted speech signal $\hat{s}(t)$ is calculated in Eq. (12). The converted speech signal $\hat{s}(t)$ was then resynthesized using the calculated mel-cepstra and the estimated f_0 of the converted source with World [16].

$$mc_i^{(\hat{s})} = mc_i^{(\hat{E})} + mc_i^{(V)} \qquad (12)$$

4. Analysis of the results

Multiple probes of training/test sets were used to train and evaluate the system. In each probe, the conversion system was trained on 40 respiration/PAL voice source recordings (45 seconds each) and was tested on 10 recordings, which were not included in the training set. The average correlation coefficient, $R(\hat{f}_0, f_0)$, between multiple probes shows a close match of 91.96±1.3% between estimated and target f_0. Figure 3.a) demonstrates a sample comparison between the estimated and target f_0. Figure 3.b) shows the corresponding respiration signals that have been used to estimate $\hat{f}_0$ together with the target PAL voice source signal ($e(t)$) from which the target f_0 has been extracted. As observed in this figure, the estimation is performed only for voiced speech denoted by a precise auto-generated voiced/unvoiced label [12]. The second GMM estimated the mel-cepstra of the PAL voice from respiration. The performance of this GMM is evaluated using the MelCD (10) between the estimated and target melcepstra. The average MelCD between all probes was 4.02±0.14 *dB*.

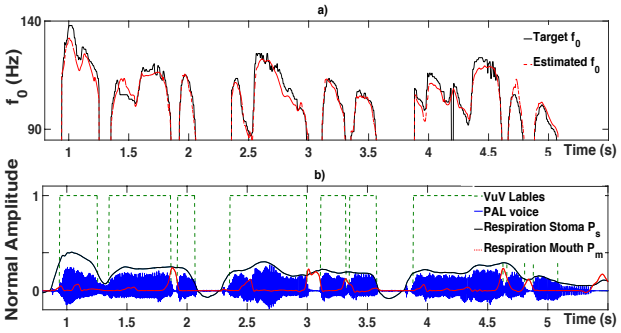

Figure 3: a) *Comparison of estimated and target f_0 using the GMM b) Underlying values of respiratory pressure p_s, p_m. The respiration recordings are time aligned with recordings of the PAL excitation source waveform (blue).*

Next the PAL speech signal was re-synthesized by substituting the estimated PAL excitation source with the original using Eq. (4). Figure 4 compares the spectrogram of the converted and original PAL speech. As observed in this figure the converted speech shows close similarities to the original.

A listening test was also performed to evaluate the performance of the system in estimating the voice source when used to re-synthesize the original PAL speech signal. Ten naïve listeners (native English speakers similar to the laryngectomy participant) were enrolled to evaluate the re-synthesized speech both in terms of quality and intelligibility. For both tests, the listeners were initially familiarized with 4 recordings of PAL speech (45 *s* each and not included in the test material). For the intelligibility test, the listeners were presented with recordings of original and re-synthesized speech in shuffled order and were asked to transcribe these. Fifty pairs of original and re-synthesized sentences were presented. The percentage of correctly identified words over the total number of words in each sentence was calculated and converted on a scale of 0 to 10 (with 10 showing maximum intelligibility). The overall Intelligibility score was averaged between all sentences for all listeners. For quality assessment, a Mean Opinion Score (MOS) was used and 25 pairs of re-synthesized and original PAL sentences were compared by the listeners, and the re-synthesized version was scored from 1 to 5 (with 5.0 being the same quality as the original PAL).

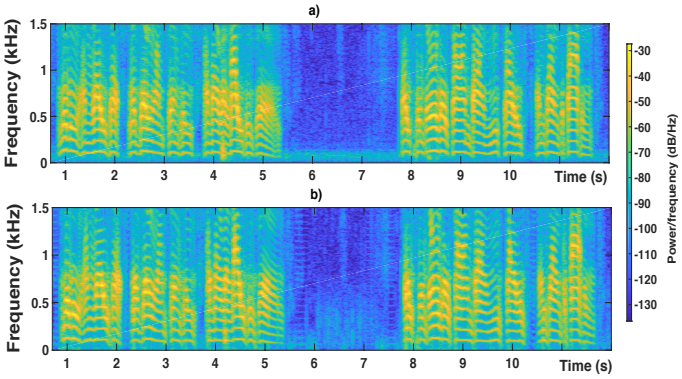

Figure 4: *The spectrum of the PAL speech signal: a) original recording using the PAL source b) Re-synthesized version using the estimated PBV source.*

The original PAL speech received an intelligibility score of 7.91±0.15 while the score for the re-synthesized version was 7.14±0.4. For quality assessment, the re-synthesized speech got a score of 3.26±0.1 compared to the original 5.0. The decay in quality was mainly attributed to the inaccuracies in substituting the mel-cepstra to re-synthesize the speech. The PAL source estimation was only for voiced phonemes. Hence, to resynthesize the unvoiced segments, we copied the mel-cepstra of the original PAL speech. The inaccuracies in the transition between estimated and copied mel-cepsta for voiced and unvoiced phonemes were perceived by listeners as a decay in the quality. However, this situation will not exist when the PBV source will be tried on the patient and the estimated source is superimposed by the natural respiration airflow of the patient to generate both voiced and unvoiced phonemes. The high intelligibility score of the PAL (7.9±0.15) compliments its strengths as being the reference of the design of Bionic Voice and is in accord with previous studies reporting similar levels of intelligibility [2, 5, 7-11]. The re-synthesized speech using the estimated PAL source also maintains less than 8% drop in the intelligibility. This, together with an accuracy of 91.96±1.3 for f_0 prediction and the MelCD of 4.02±0.14 *dB* between estimated and target mel-cepstra, advocates the strengths of the proposed method to estimate the PAL excitation waveform from the respiration.

5. Conclusions

This study defines a novel framework which effectively estimates the excitation waveform of a Pneumatic Artificial Larynx (PAL) from respiration signals. Such framework will be employed in designing Pneumatic Bionic Voice (PBV) sources [12]. The PBV has tremendous potential to replace the existing gold standard of TE voice prostheses. Mimicking the PAL excitation, provides the PBV with a much wider range of f_0 patterns. The PBV is also expected to step beyond the high quality of the PAL and sound more natural by applying VC frameworks to paired datasets of PAL and natural speech.

6. Acknowledgements

The authors wish to thank Mr. Ben Binyamin and Mr. Colin Symons from the MARCS Institute, Western Sydney University, for their assistance in data collection of this study.

Part of this work was supported by JST, PRESTO Grant Number JPMJPR1657.

7. References

[1] K. Delsupehe, I. Zink, M. Lejaegere, and P. Delaere, "Prospective randomized comparative study of tracheoesophageal voice prosthesis: Blom-singer versus provox," *The Laryngoscope,* vol. 108, no. 10, pp. 1561-1565, 1998.

[2] B. Weinberg and A. Riekena, "Speech produced with the Tokyo artificial larynx," *Journal of Speech and Hearing Disorders,* vol. 38, no. 3, pp. 383-389, 1973.

[3] G. S. Meltzner and R. E. Hillman, "Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech," *Journal of Speech, Language and Hearing Research,* vol. 48, no. 4, p. 766, 2005.

[4] A. K. Fuchs, M. Hagmüller, and G. Kubin, "The New Bionic Electro-Larynx Speech System," *IEEE Journal of Selected Topics in Signal Processing,* vol. 10, no. 5, pp. 952-961, 2016.

[5] R. L. Goode, "Artificial laryngeal devices in post-laryngectomy rehabilitation," *The Laryngoscope,* vol. 85, no. 4, pp. 677-689, 1975.

[6] J. J. Xu, X. Chen, M. P. Lu, and M. Z. Qiao, "Perceptual evaluation and acoustic analysis of pneumatic artificial larynx," *Otolaryngology and Head and Neck Surgery,* vol. 141, no. 6, pp. 776-780, 2009.

[7] M. L. Ng, C.-L. I. Kwok, and S.-F. W. Chow, "Speech performance of adult Cantonese-speaking laryngectomees using different types of alaryngeal phonation," *Journal of Voice,* vol. 11, no. 3, pp. 338-344, 1997.

[8] S. Bennett and B. Weinberg, "Acceptability ratings of normal, esophageal, and artificial larynx speech," *Journal of Speech, Language and Hearing Research,* vol. 16, no. 4, p. 608, 1973.

[9] I. K.-Y. Law, E. P.-M. Ma, and E. M.-L. Yiu, "Speech intelligibility, acceptability, and communication-related quality of life in Chinese alaryngeal speakers," *Archives of Otolaryngology - Head and Neck Surgery,* vol. 135, no. 7, p. 704, 2009.

[10] T. Y. Ching, R. Williams, and A. V. Hasselt, "Communication of lexical tones in Cantonese alaryngeal speech," *Journal of Speech, Language and Hearing Research,* vol. 37, no. 3, p. 557, 1994.

[11] S. Singer *et al.*, "Speech rehabilitation during the first year after total laryngectomy," *Head and neck,* pp. 1-8, 2012.

[12] F. Ahmadi, F. Noorian, D. Novakovic, and A. van Schaik, "A pneumatic Bionic Voice prosthesis—Pre-clinical trials of controlling the voice onset and offset," *PloS one,* vol. 13, no. 2, p. e0192257, 2018.

[13] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 9, pp. 2505-2517, 2012.

[14] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[15] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Technical Report ICSI-TR-97-02, University of Berkeley1997.

[16] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems,* vol. 99, no. 7, pp. 1877-1884, 2016.

[17] M. Zecca, S. Micera, M. Carrozza, and P. Dario, "Control of multifunctional prosthetic hands by processing the electromyographic signal," *Critical Reviews™ in Biomedical Engineering,* vol. 30, no. 4-6, 2002.

[18] M. Zardoshti-Kermani, B. C. Wheeler, K. Badie, and R. M. Hashemi, "EMG feature evaluation for movement control of upper extremity prostheses," *IEEE Transactions on Rehabilitation Engineering,* vol. 3, no. 4, pp. 324-333, 1995.

[19] R. Boostani and M. H. Moradi, "Evaluation of the forearm EMG signal features for the control of a prosthetic hand," *Physiological measurement,* vol. 24, no. 2, p. 309, 2003.

[20] F. Ahmadi, M. A. Ribeiro, and M. Halaki, "Surface electromyography of neck strap muscles for estimating the intended pitch of a bionic voice source," in *Biomedical Circuits and Systems Conference (BioCAS), 2014 IEEE*, 2014, pp. 37-40: IEEE.

[21] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing* vol. 15, no. 8, pp. 2222-2235, 2007.

[22] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," *Github:* https://github.com/google/REAPER, 2015.

[23] W. De Armas, "Vocal Frequency Estimation and Voicing State Prediction with Surface EMG Pattern Recognition," MSc Thesis, University of Toronto, 2013.

[24] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 1, pp. 24-33, 2007.

[25] H. Takahashi, M. Nakao, Y. Kikuchi, and K. Kaga, "Intra-Oral Pressure–Based Voicing Control of Electrolaryngeal Speech with Intra-Oral Vibrator," *Journal of Voice,* vol. 22, no. 4, pp. 420-429, 2008.