INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency

Tuan Dinh[1], *Alexander Kain*[1], *Robin Samlan*[2], *Beiming Cao*[3], *Jun Wang*[3]

[1]Oregon Health & Science University
[2]University of Arizona
[3]University of Texas at Austin

dintu@ohsu.edu, kaina@ohsu.edu, rsamlan@arizona.edu, beiming.cao@utexas.edu, jun.wang@austin.utexas.edu

Abstract

Individuals who undergo a laryngectomy lose their ability to phonate. Yet current treatment options allow alaryngeal speech, they struggle in their daily communication and social life due to the low intelligibility of their speech. In this paper, we presented two conversion methods for increasing intelligibility and naturalness of speech produced by laryngectomees (LAR). The first method used a deep neural network for predicting binary voicing/unvoicing or the degree of aperiodicity. The second method used a conditional generative adversarial network to learn the mapping from LAR speech spectra to clearly-articulated speech spectra. We also created a synthetic fundamental frequency trajectory with an intonation model consisting of phrase and accent curves. For the two conversion methods, we showed that adaptation always increased the performance of pre-trained models, objectively. In subjective testing involving four LAR speakers, we significantly improved the naturalness of two speakers, and we also significantly improved the intelligibility of one speaker.
Index Terms: speech intelligibility, voice conversion, total laryngectomy

1. Introduction

Speech is arguably the most important biosignal for human communication. Pressure from the lungs drives typical laryngeal voice and speech. The pharynx, tongue, and lips shape exhaled air to produce voiceless sounds, and quasiperiodic vocal fold vibration creates the sound wave that vocal tract constrictions shape into vowels and voiced consonants. Individuals who undergo a laryngectomy lose their ability to produce speech sounds normally because their trachea is disconnected from the vocal tract. Laryngectomy is performed as surgical treatment for advanced laryngeal and hypopharyngeal cancers. These patients experience a lower quality of life because of their atypical speech (we term this as LAR speech) during social interactions, as they believe that other people perceive them as abnormal, or they directly experience symbolic violence [1]. In 2020, an estimated 12,370 new cases of laryngeal cancers are expected in the U. S. [2]. Although the incidence of laryngeal cancers is decreasing due to the decreasing number of smokers, there is still a large projected number in the next decades because the rate of decrease is only 2–3% [2].

There are currently a limited number of alternative communication options for laryngectomees. Typing-based alternative and augmentative communication (AAC) devices are slow and limited by the speed of typing. The main speech options for individuals after laryngectomy are (1) *esophageal* speech (pushing air from the mouth to the pharyngo-esophageal segment (PES) and using the PES for vibration), (2) tracheo-esophageal puncture (TEP) speech wherein speakers use lung air to power PES vibration for voiced speech, and (3) use of an artificial larynx, in the form of either an external electrolarynx placed on the neck (ELX) or with an intraoral tube. The ELX generates an electronic sound source that is shaped by the lips and tongue into always-voiced speech at a constant pitch [3]. These options are suboptimal: esophageal speech requires extensive training and practice and is difficult to learn [4], TEP is a surgical operation and requires talkers to place their thumb over their stoma during the speech act, which is rarely hands-free and poses certain risks, and the artificial larynx produces a very robotic-like sounding voice. All options produce unnatural sounding and difficult to understand speech for several reasons, including poor voice quality, voiced/voiceless differentiation, and articulatory precision [5, 6].

Voice conversion (VC) can be used to alter and improve LAR speech. In this paper, we aim to increase the intelligibility and naturalness of LAR speech, by applying three innovative methods: (1) predicting binary voicing/unvoicing and the degree of aperiodicity parameters, as used by the WORLD vocoder [7, 8], (2) improving the spectral characteristics by mapping from LAR to clearly-articulated speech (CLR), and (3) creating a suitable F0 trajectory (see Figure 1).

2. Related Work

For two decades, researchers have attempted to create natural-sounding speech for laryngectomees. Using rule-based spectral VC approaches, some differences between alaryngeal speech and normal speech can be compensated for by modifying speech formant properties. For example, in an early work, the authors made esophageal speech more intelligible by expanding the formant bandwidths [9]. Another approach is to decrease formants' frequencies using formant shifting methods [10, 11], since it was found that speakers who underwent partial laryngectomy shift their formants to higher frequencies due to the shortened vocal tract length [12]. Similarly, TEP speech is reported to have a spectral tilt which favors the high frequency band [13]; thus, the authors used a 6 dB/octave roll-off filter to de-emphasize the high frequency band. These approaches led to limited improvement in intelligibility or naturalness.

For statistical VC, previous approaches included the use of Gaussian mixture models [14, 15] and deep neural networks [16, 17] for mapping spectral features. These models are limited because of over-smoothing of converted spectra, leading to muffled speech [18, 19]. Recently, generative adversarial networks (GANs) [20] have been shown effectively address the over-smoothing problem in VC [19] and speech synthesis [21].

This material is based upon work supported by the National Institutes of Health under grants R01DC016621 and R03DC013990.

http://dx.doi.org/10.21437/Interspeech.2020-1196

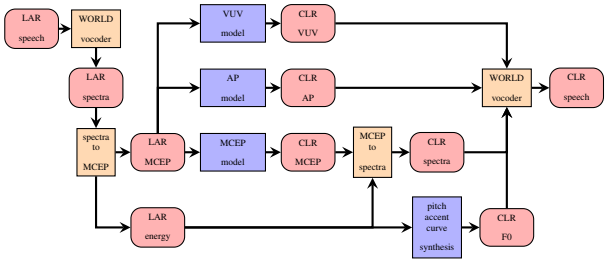

Figure 1: *Flowchart of approach during prediction*

The LAR-to-CLR spectral mapping can be viewed as an image-to-image translation task, in which the image is a window of the time-frequency representation of speech. In image-to-image translation, a conditional GAN (cGAN) [22] proved to be effective in generating less blurry images by combining a traditional adversarial loss and a mean absolute reconstruction loss (or L1 loss). In this paper, we leverage the cGAN architecture for mapping LAR features to CLR features.

Generally, LAR speech lacks reliable voicing and fundamental frequency (F0) information, but also meaningful F0 variability. One approach used to produce a more natural sounding speech signal is therefore to create converted or synthetic voicing and F0 trajectories. In related work on reconstructing normal speech from whispered speech, F0 values were estimated from filtered gain parameters [23] or using the first formant frequency and its magnitude [24].

For F0 conversion or synthesis, a variety of approaches have been proposed. The introduction of jitter (small pitch perturbations) was found to reduce the artificiality of ELX speech [25]. Another approach created an artificial pitch contour by means of filtering, scaling and offsetting the energy envelope [10]. Alternatively, the F0 trajectory can be generated using formant frequencies as well as gains from a linear prediction model [26]. In a third approach, TEP speech is first converted to whispered speech, and then an F0 trajectory was synthesized by adding a normalized short-term energy contour of the whispered speech to an average pitch when the energy value is greater than a threshold [27]. In the paper, we directly use the normalized short-term energy contour of LAR speech in combination with a simple intonation model to synthesize the final F0 trajectory.

3. Data

For the *source* LAR speech, we used a database of 4 male speakers consisting of 3 LAR-TEP speakers (L001, L002, L006) and 1 LAR-ELX speaker (L004); the average age was 61.75±8.77. The speakers underwent total laryngectomy. The pitch of the LAR-TEP speech is low and highly-variable, and voicing correlates largely with energy. Speech analysis using standard voicing and fundamental frequency (F0) analysis algorithms fail for this type of speech the majority of the time. The LAR-ELX speech is always voiced with a constant F0 (80 Hz in our case). All speakers read all sentences of the AAC132 list [28]. For the *target* CLR speech, we created a synthetic male voice using Tacotron 2 [29] with the Waveglow vocoder [30]. We created utterances to match those of the source. The database was divided into 100/16/16 sentences for training/validation/testing. All waveforms were resampled from 22.05 kHz to 16 kHz.

We also used a multi-speaker TIMIT database [31] for pre-training. Moreover, to simulate the characteristics of LAR-TEP and LAR-ELX speech, we also created a fully-unvoiced (FU) TIMIT and a fully-voiced (FV) TIMIT. We used a process of first analyzing standard normally-voiced (NV) TIMIT using the WORLD vocoder [7, 8], then setting all frames to either unvoiced, or voiced with all F0 values set at a constant value of 80 Hz to create FU-TIMIT and FV-TIMIT, respectively. Of the 630 available speakers we used all 462/144/24 speakers designated for training, validation, and testing, respectively. By convention, we eliminated the spoken dialect samples (SA sentences) for all speakers.

4. Predicting Voicing and Degree of Voicing

In the section, we investigated the capability of predicting voicing presence and the degree of voicing. Specifically, we predicted a binary voicing value and continuous 2-band aperiodicity [8] values from mel-cepstral coefficients (MCEP), using deep neural networks (DNN). We used 32$^{\text{nd}}$-order MCEP (MCEP-32) features to avoid pitch effects in the spectral representation. We pre-trained three speaker-independent DNNs using normal, FU-, and FV-TIMIT. Then we adapted all three models on LAR-TEP and LAR-ELX speech, to examine whether the choice of pre-training database is important for the two different types of LAR speech. Our motivation was based on the hypothesis that FU-TIMIT is more similar to the unvoiced nature of LAR-TEP speech, and FV-TIMIT is more similar to the always voiced LAR-ELX speech (we use the terms voiced and unvoiced from a signal analysis point of view, not a production point of view). In addition, we also investigated how performance is affected by different context window lengths.

4.1. Pre-training

We analyzed the NV/FU/FV-TIMIT databases using the WORLD vocoder to obtain F0, voicing, spectrogram, and 2-band aperiodicity parameters, for each frame, using a frame rate of 5 ms. The voicing is a binary voiced/unvoiced flag (VUV). The 2-band aperiodicity (AP) is a single scalar representing the degree of voicing at 3000 Hz, which is the boundary frequency of the two frequency bands: $[0, 3000]$ and $[3000, 8000]$ Hz. We extracted MCEP-32 from the WORLD spectrogram. We excluded the zeroth coefficient (representing energy), then trained DNNs to predict NV-TIMIT VUV or AP parameters from each of the three NV/FU/FV-TIMIT MCEP coefficients 1–31. We normalized inputs of the network with standard scaling. We added context by concatenating the current frame with preceding and following frames. We considered context lengths of 25, 55, and 105 ms, respectively.

The DNN has three hidden layers with 256 nodes each. The activation function is parametric ReLU. Each hidden layer is preceded by batch normalization (except the first layer), and followed by dropout with a dropout rate of 0.2 (except the last layer). We trained using the Adam optimizer, a mini-batch size of 256, and early stopping. The binary cross entropy and mean-squared error loss functions were used for voicing classification and 2-band aperiodicity regression, respectively. In total, there are 18 (3 training databases $\times$ 3 context lengths $\times$ 2 output types) pre-trained models.

We objectively evaluated the performance of each model using balanced accuracy (BAC, defined as average recall) for VUV classification (since the classes were imbalanced), and r^2 for AP regression. On average, we obtained a BAC/r^2 of 0.94/0.75. The results suggest that we can predict voicing and the degree of voicing from spectral shape alone. The different context lengths did not result in significant differences in

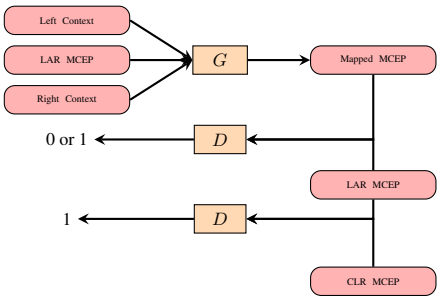

Figure 2: *cGAN framework for style conversion*

BAC and r^2. We then tested the pre-trained networks without any adaptation to predict target VUV or AP from LAR-TEP and LAR-ELX speech with 16 test sentences. We can see that the BAC and r^2 drastically decrease. We have an approximate BAC/r^2 of 0.60/$-$0.44 for L001, 0.58/$-$0.68 for L002, 0.49/$-$0.4 for L004, 0.48/$-$0.6 for L006.

4.2. Adaptation

We adapted the pre-trained networks to individual speakers' LAR-TEP or LAR-ELX speech to improve performance. We aligned each LAR utterance to its parallel target CLR utterance using dynamic time warping on 32$^{\text{nd}}$-order log filter bank features. There are 72 (18 pre-trained models $\times$ 4 speakers) adapted models. All training settings were the same as those of pre-training. Because there were many more voiced frames than unvoiced frames, we over-sampled the unvoiced frames to balanced the classes. The average r^2 of AP and the BAC of VUV were ˜0.24/˜0.7 for L001, ˜0.44/0.73 for L002, ˜0.28/0.70 for L004, and ˜0.05/0.65 for L006; it appears that some speakers' AP was much easier to predict than others', whereas VUV prediction performance was similar. As expected, adaptation always improved performance. Varying context size resulted in a relatively narrow BAC range from 0.65 to 0.73, and thus we used 55 ms from this point forward. Surprisingly, pre-training with FU- and FV-TIMIT as opposed to NV-TIMIT did not show improved performance.

5. Predicting Spectrum

5.1. Conditional Generative Adversarial Network

Traditional GANs have a generative model or a generator (G) and a discriminative model or a discriminator (D), that together play a min-max game. Component G tries to fool component D by generating outputs similar to the real data, while component D is trained to distinguish the output of component G from real data. Component G is a mapping function from random noise z to y, $G : \{z\} \rightarrow y$ [20]. In contrast, a cGAN model learns a mapping from an input x and random noise z to y, $G : \{x, z\} \rightarrow y$. The cGAN model has both G and D conditioned on input x [22], trained with the objective function $\mathcal{L}(D, G)$:

$$\min_G \max_D \mathcal{L}(D, G) = \quad (1)$$
$$\mathbb{E}_{x,y}\left[\log D(x, y)\right] + \mathbb{E}_{x,z}\left[\log(1 - D(x, G(x, z)))\right]$$

In our cGAN, we did not use random noise z because it has proven ineffective for generator G [22, 32]. Instead, our generator mapped LAR speech features to aligned CLR speech features as shown in Figure 2. For the input vector of G, we added context by concatenating the current LAR MCEP-32 frame with five preceding and five following frames. We normalized the inputs and outputs of the network via standard scaling. The input of D consisted of a single frame of either the output of G or an aligned CLR feature frame, combined with the current LAR feature frame (what we wanted the output to be conditioned on). Thus, both G and D are conditioned on the current LAR feature frame. In addition to the adversarial loss function $\mathcal{L}(D, G)$ in Equation 1, we also minimized the L1 loss between the output of G and the ground truth; this addition was demonstrated to generate less blurry output compared to a root-mean-squared reconstruction loss in an image task [22]. We added the L1 loss with a weighting factor of 100 to $\mathcal{L}(D, G)$.

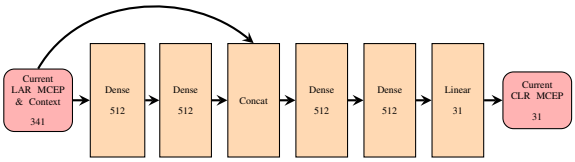

Figure 3: *Generator architecture*

The structure of the generator G, shown in Figure 3, is similar to our previous work [33]; however, there is no skip connection which adds the input of G to the output of its final dense layer, because performance worsened when using the skip connection. The discriminator D is a DNN with two hidden layers with 256 nodes each, and a single-node output layer with sigmoidal activation function. To help stabilize the training process, we used (1) a leaky ReLU activation function with a slope of 0.2 for negative inputs for both G and D, (2) a dropout layer following each hidden layer of D with a dropout rate of 0.5, (3) the Adam optimizer with a batch size of 128, and (4) weights initialized from a zero-centered normal distribution with standard deviation 0.02 [34]. We used a momentum of 0.5, a learning rate decay of 10^{-5}, and learning rate of 10^{-4} for D, and $2 \cdot 10^{-4}$ for G.

5.2. Predicting spectrum

We first pre-trained the cGAN with the methods described above to convert FU- and FV-TIMIT MCEPs to NV-TIMIT MCEPs, excluding the zeroth (energy) coefficient (similar to 4.1). This was a data-rich proxy for the eventual mapping of LAR to CLR speech. There were two pre-trained models, one for each training databases. We passed the source energy unmodified to the target features. We then calculated the predicted spectra and compared them to target spectra in terms of log spectral distortion (LSD). On average, the LSDs were 7.64 and 6.46 dB for FU- and FV-TIMIT, respectively. We then tested the pre-trained models on LAR speech, obtaining 60 dB for L001, 45 dB for L002, 51 dB for L004, and 62 dB for L006. We then adapted the cGAN conversion to convert LAR MCEP to CLR MCEP (similar to 4.2). There were eight adapted models (2 pre-trained models $\times$ 4 speakers). We adapted in two ways: adapting only the generator or adapting both the generator and the discriminator. We observed that the latter yields lower LSD scores. The average LSD was 32 dB for L001, 33 dB for L002, 31.6 dB for L004, and 37.4 dB for L006. As expected, the adaptation always increased performance. Finally, pre-training with FU- and FV-TIMIT did not improve the results.

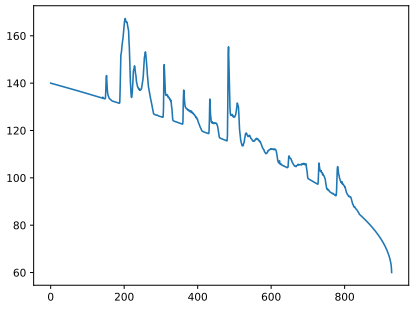

Figure 4: *Example synthetic F0 trajectory*

6. Synthesis

For analysis and synthesis, we used the WORLD vocoder [7, 8]. We analyzed LAR speech into F0, VUV, AP, and pitch-synchronous spectrogram, from which we derived MCEP-32 features. We predicted CLR VUV and AP using the DNN (see Section 4), and CLR MCEPs using the cGAN (see Section 5). The source energy was passed through unmodified. We synthesized the CLR F0 from the source energy, using a method described below. The predicted/synthetic set of CLR parameters was used to synthesize the final speech waveform.

6.1. Synthetic pitch accent curve

We used a simple model of intonation consisting of phrase and accent curves [35]. The phrase curve p is defined as

$$p(t) = p_{\text{min}} + (p_{\text{max}} - p_{\text{min}}) \left(1 - \frac{t}{T}\right)^{b}$$

where we empirically set p_{max}=140, p_{min}=60, and b=0.5; t is a time index between 0 and T. To set accent curve α, we used $a(t) = A \cdot e(t)$ where we empirically set A=40, and e is the max-normalized energy. The final F0 trajectory is calculated as $f_0(t) = p(t) + a(t)$. Figure 4 shows an example of the synthetic F0 trajectory. Informal perceptual experiments confirmed that replacing a natural F0 trajectory with a synthetic one did not reduce the naturalness of speech.

7. Evaluation

We evaluated the efficacy of using predicted VUV, AP, F0, and spectrum in LAR speech in term of naturalness and intelligibility in two comparative mean opinion score (CMOS) tests. The LAR denotes vocoded LAR speech. Informal listening tests have shown that vocoded LAR speech is sufficiently close to the original LAR speech in terms of intelligibility and naturalness. The CLR-spectrum denotes LAR speech with predicted MCEPs. The CLR-intonation denotes LAR speech with predicted VUV, AP and F0. The CLR-all denotes predicting all vocoder parameters except energy. We compared LAR to all other conditions. Each of the two CMOS tests consisted of 16 sentences $\times$ 4 speakers $\times$ 3 pairs of conditions = 192 unique trials. We limited each listener to hear each unique sentence once (presentation order was randomized); therefore we needed blocks of 192 $\div$ 16 = 12 listeners to cover all trials. Both experiments were conducted on Amazon Mechanical Turk (AMT); we required listeners to have an approval rate $\geq$ 90% and to live in the U. S. Each test had 48 listeners, for a total of 96 listeners. In each trial, participants listened to samples A and B in sequence and were then asked: "Is A more natural than B?" or "Is A more intelligible than B?" for the naturalness and intelligibility tests, respectively. Responses were selected from a 5-point scale that consisted of "definitely better" (+2), "better" (+1), "same" (0), "worse" (−1), and "definitely worse" (−2).

Table 1: *Perceptual CMOS test results comparing modified conditions against the vocoded LAR speech condition. CLR-spectrum, CLR-intonation, CLR-all denote predicting CLR spectrum, CLR VUV/AP/F0, or a combination of these, respectively. Scores marked with an asterisk are significantly different.*

naturalness	CLR-spectrum	CLR-intonation	CLR-all
L001 (TEP)	−0.0	−0.3*	*0.4**
L002 (TEP)	−0.1	−0.0	0.1
L004 (ELX)	−0.56*	−0.25	0.22
L006 (TEP)	−0.3*	−0.2*	*0.7**

intelligibility	CLR-spectrum	CLR-intonation	CLR-all
L001 (TEP)	−0.1	−0.1	0.1
L002 (TEP)	0.1	0.2	−0.3*
L004 (ELX)	−0.34*	*0.34**	−0.2
L006 (TEP)	0.2	−0.1	−0.0

Table 1 top shows the pair-wise relative naturalness of the stimuli. Positive scores show improvement over LAR speech. Converting VUV, AP, and MCEPs, and using synthetic F0, improved the naturalness of LAR-TEP and LAR-ELX, statistically significantly ($p < 0.01$) as compared to zero (no preference) in a one-sample t-test. Table 1 bottom shows the pair-wise relative intelligibility of the stimuli. We improved intelligibility statistically significantly ($p < 0.01$, as compared to zero in a one-sample t-test) only for L004 (LAR-ELX speech) when predicting VUV, AP, and using a synthetic F0. This is probably because there was no sufficient voicing information in TEP speech. Further studies with a larger number of patients will be conducted to verify these preliminary findings.

8. Conclusion

We proposed two conversion methods to improve naturalness and intelligibility of LAR speech: 1) predicting CLR VUV or CLR AP using a DNN, and 2) predicting CLR MCEPs using a cGANs. We also created a synthetic F0 trajectory with an intonation model consisting of phrase and accent curves. For predicting CLR VUV or CLR AP, using different context lengths did not have a significant impact. Moreover, pre-training the prediction networks on FU-, and FV-TIMIT as opposed to NV-TIMIT did not result in improved performance. Similarly, pre-training with FU-, and FV-TIMIT did not lead to improved performance for predicting CLR MCEP. Adaptation always improved performance. In our subjective tests with four LAR speakers, we significantly improved the naturalness of two speakers, and we significantly improved the intelligibility of one speaker. The results are promising for a challenging task with a lot of individual variability among four LAR speakers.

9. References

[1] J. Mertl, E. Žáčková, and B. Řepová, "Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 4, pp. 342–352, 2018.

[2] American cancer society 2020. [Online]. Available: https://www.cancer.org/cancer/laryngeal-and-hypopharyngeal-cancer/about/key-statistics.html [retrieved on 05-15-2020]

[3] H. Liu and M. L. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[4] B. J. Bailey, J. T. Johnson, and S. D. Newlands, *Head & Neck Surgery–Otolaryngology*. Lippincott Williams & Wilkins, 2006, vol. 1.

[5] N. Bi and Q. Yingyong, "Speech conversion and its application to alaryngeal speech enhancement," in *ICSP*, 1996, pp. 1586–1589.

[6] Z. A. Khan, P. Green, S. Creer, and S. Cunningham, "Reconstructing the voice of an individual following laryngectomy," *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–661, 2011.

[7] 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. E99-D, no. 7, pp. 1877–1884, 2016.

[8] M. Morise, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[9] R. H. Ali and S. B. Jebara, "Esophageal speech enhancement using excitation source synthesis and formant patterns modification," in *Signal-Image Technology & Internet Based Systems*, 2006, pp. 615–624.

[10] A. Loscos and J. Bonada, "Esophageal voice enhancement by modeling radiated pulses in frequency domain," in *121st Convention of the Audio Engineering Society*, 2006.

[11] S. H. R. annd Ian 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.

[12] R. A. Kazi, V. M. Prasad, J. Kanagalingam, C. M. Nutting, P. Clarke, P. Rhys-Evans, and K. J. Harrington, "Assessment of the formant frequencies in normal and laryngectomized individuals using linear predictive coding," *Journal of Voice*, vol. 21, no. 6, pp. 661–668, 2007.

[13] A. D. Pozo and S. Young, "Continuous tracheosophageal speech repair," in *the European Signal Processing Conference*, 2006.

[14] N. Keigo, T. Toda, H. Saruwatari, and K. Shikano, "Electrolaryngeal speech enhancement based on statistical voice conversion," in *INTERSPEECH*, 2009, pp. 1431–1434.

[15] ——, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[16] K. Kazuhiro and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on cldnn," in *EUSIPCO*, 2018, pp. 2115–2119.

[17] O. I. Ben, J. D. Martino, and K. Ouni, "Enhancement of esophageal speech obtained by a voice conversion technique using time dilated fourier cepstra," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 99–110, 2019.

[18] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *ICASSP*, 2005.

[19] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH*, 2017.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[21] N. H. Takuhiro Kaneko, Hirokazu Kameoka, Y. Ijima, K. Hiramatsu, and K. Kashiro, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *ICASSP*, 2017.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2016, pp. 5967–5976.

[23] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7–8, pp. 515–520, 2002.

[24] McLoughlin, I. Vince, J. Li, and Y. Song, "Reconstruction of continuous voiced speech from whispers," in *INTERSPEECH*, 2013.

[25] K. Chenausky and J. MacAuslan, "Utilization of microprocessors in voice quality improvement: the electrolarynx," *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 8, no. 3, pp. 138–142, 2000.

[26] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Regeneration of speech in voice-loss patients," in *13th International Conference on Biomedical Engineering*, 2009, pp. 1065–1068.

[27] A. R. N, A. Rao MV, G. N. Meenakshi, and P. K. Ghosh, "Reconstructing neutral speech from tracheoesophageal speech," in *INTERSPEECH*, 2018, pp. 1541–1545.

[28] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang, "Permanent magnetic articulograph (pma) vs electromagnetic articulograph (ema) in articulation-to-speech synthesis for silent speech interface," in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 2019, pp. 17–23.

[29] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[30] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP*, 2019.

[31] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *ICASSP*, 2006.

[32] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, 2017.

[33] T. Dinh, A. Kain, and K. Tjaden, "Using a manifold vocoder for spectral voice and style conversion," in *INTERSPEECH*, 2019, pp. 1388–1392.

[34] S. Chintala, "How to train a gan?" https://github.com/soumith/ganhacks, 2016.

[35] J. P. van Santen, T. Mishra, and E. Klabbers, "Estimating phrase curves in the general superpositional intonation model," in *Fifth ISCAA Workshop on Speech Synthesis*, 2004.