Alaryngeal Speech Generation Using MaskCycleGAN-VC and Timbre-Enhanced Loss

Hnin Yadana Lwin
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
hnin.lwin@mail.kmutt.ac.th

Wuttipong Kumwilaisak
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
wuttipong.kum@kmutt.ac.th

Chatchawarn Hansakunbuntheung
National Science and Technology Development Agency
Pathum Thani, Thailand
chatchawarn.han@nstda.or.th

Nattanun ThatphiThakkul
National Science and Technology Development Agency
Pathum Thani, Thailand
nattanun.tha@nstda.or.th

ABSTRACT

This paper introduces a data augmentation technique for alaryngeal speech using voice conversion within the MaskCycleGAN-VC framework [6]. Our method leverages two masking techniques: Articulatory Dimension Masking (ADM) and the combination of ADM with Consecutive Time Masking (CTM), called SpecAugment[11]. The initial technique used for masking within the MaskCycleGAN-VC framework is CTM, and our proposed additional masking techniques enhance the quality and performance of voice conversion for alaryngeal speech. We can also expand the variability of voice characteristics within the converted alaryngeal speech dataset. One notable enhancement in our approach is incorporating a timbre similarity score into the generator loss, known as the Timbre Enhanced Loss. This score dynamically guides the conversion process to prioritize preserving timbral characteristics during voice transformation. From our experiments using different objective metrics, the proposed method can provide synthesized alaryngeal speeches having characteristics close to the actual ones.

CCS CONCEPTS

• **Computing methodologies** → **Voice Conversion**.

KEYWORDS

Data Augmentation, Voice Conversion, Consecutive Time Masking, Articulatory Dimension Masking, SpecAugment, Timbre Enhanced Loss

ACM Reference Format:
Hnin Yadana Lwin, Wuttipong Kumwilaisak, Chatchawarn Hansakunbuntheung, and Nattanun ThatphiThakkul. 2023. Alaryngeal Speech Generation Using MaskCycleGAN-VC and Timbre-Enhanced Loss. In *13th International Conference on Advances in Information Technology (IAIT 2023), December 06–09, 2023, Bangkok, Thailand*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3628454.3631582

IAIT 2023, December 06–09, 2023, Bangkok, Thailand

ACM ISBN 979-8-4007-0849-7/23/12.
https://doi.org/10.1145/3628454.3631582

1 INTRODUCTION

Alaryngeal speech refers to speech produced by individuals who have had their larynx removed, typically due to medical conditions like laryngeal cancer. In general, alaryngeal speech data is relatively rare, primarily due to the unique nature of alaryngeal speech and the challenges it presents. Data from such individuals is limited, and collecting a diverse dataset of alaryngeal speech can be difficult. Moreover, working with individuals who have undergone a laryngectomy requires ethical considerations and obtaining informed consent. As a result, the research on augmenting alaryngeal speech data that solves the limitations and ethics of human subjects should be prioritized.

Many studies have been related to the characteristics and perceptual evaluation of alaryngeal[13] . Marinela et al.[13] assessed and compared the self-reported vocal limitations experienced by laryngectomees who underwent three distinct communication approaches: tracheoesophageal speech, esophageal speech, and electrolarynx. Recently, Cao et al.[3] has demonstrated a practical approach for data augmentation in End-to-End Speech Recognition for laryngectomees. However, prior works have not yet to explore data augmentation through voice conversion. Their primary focus, as highlighted in their work[3], centered exclusively on silent speech recognition without placing significant emphasis on the quality or effectiveness of the augmented data used in their research. In voice conversion, models like CycleGAN-VC2[4] and CycleGAN-VC3[5] have emerged. While CycleGAN-VC2 struggles with capturing time-frequency structures, CycleGAN-VC3 improves by introducing a time-frequency adaptive normalization (TFAN) module. However, this approach increases the number of parameters to be learned. To address the problem in CycleGAN-VC3, MaskCycleGAN-VC introduces FIF(Filling in Frame), in which we utilize a temporal mask on the input mel-spectrogram, promoting the converter to complete absent frames by considering information from neighboring frames. For object evaluation, two metrics were used in CycleGAN-VC2. Firstly, they utilized the Mel-cepstral distortion (MCD) to gauge the disparity between the converted and target MCEP sequences. Secondly, to assess local structural distinctions, they employed the modulation spectra distance (MSD), which quantifies the root mean square error between the logarithmic modulation spectra of MCEPs for both the target and converted data, averaged across all MCEP dimensions and modulation frequencies. In MaskCycleGAN-VC[6], they replaced MSD with the Kernel DeepSpeech Distance

(KDSD)[2], a method that calculates the highest average divergence within the feature space of DeepSpeech2[1]. This metric has been demonstrated to have a strong correlation with human judgement.

In this paper, we utilize voice conversion to generate alaryngeal speech, as there has yet to be prior research for alaryngeal speech generation using VC. We used MaskCycleGAN-VC framework [3]as our baseline model as it is the latest voice conversion method that outperforms CycleGAN-VC2 and CycleGAN-VC3. Our method leverages two innovative masking techniques: Articulatory Dimension Masking (ADM) and the combination of ADM with Consecutive Time Masking (CTM), called SpecAugment. The initial approach for incorporating masking within the MaskCycleGAN-VC framework involved using a technique known as CTM. However, our innovative masking techniques go beyond CTM and offer a broader range of options. These novel masking techniques significantly diversify the dataset, enhancing the variability of voice characteristics in the converted alaryngeal speech dataset. This expanded variability is crucial for achieving more comprehensive and accurate results in some applications such as ASR(Automatic Speech Recognition) or SSR(Silent Speech Recognition). One notable enhancement in our approach is incorporating a timbre similarity score into the generator loss, known as the Timbre Enhanced Loss. This score dynamically guides the conversion process to prioritize preserving timbral characteristics during voice transformation. We utilize MCD(Mel Cepstral Distortion), FAD(Frechet Audio Distance, Timbre Similarity, and F0 comparison for objective evaluation.

This paper is organized as follows. Section 2 describes the MaskCycleGAN-VC with different masking methods used to generate alaryngeal speeches. Section 3 explains the timbre-enhanced loss that is integrated to the cost function of the MaskCycleGAN-VC. The experimental results are discussed in Section 4. Finally, conclusion remarks are in Section 4.

2 ALARYNGEAL SPEECH CONVERSION WITH MASKCYCLEGAN-VC

The MaskCycleGAN-VC is an extension of the CycleGAN-VC2 [4]. The CycleGAN-VC2 trains a voice converter G that converts source acoustic features to target acoustic features. The training process relies on an adversarial loss [4], cycle-consistency loss [4], second adversarial loss [4], and identify mapping loss [4]. The total loss function used to train the CycleGAN-VC can be expressed as

$$L_{total} = L_adv + \lambda_{cyc} L_{cyc} + \lambda_{id} L_{id} + L_{adv2}, \tag{1}$$

where L_adv, L_{cyc}, L_{id}, and L_{adv2} are an adversarial loss, cycle-consistency loss, identify mapping loss, and second adversarial loss, respectively.

We utilize the MaskCycleGAN[3] to generate alaryngeal speeches from regular speeches. The MaskCycleGAN-VC is the latest speech conversion method employing the filling-in-frames (FIF) technique. It applies a mask to the input Mel-spectrogram. Then, the voice converter will fill in the missing frames using knowledge from neighboring frames. As a result, the voice converter can learn time-frequency structure in a self-supervised manner. During the training, let the mel-spectrogram of the regular speech be x. We introduce a mask to the mel-spectrogram. Apart from the original consecutive temporal masking (CTM) in [3], we add more alternative masks, including articulatory dimension masking (ADM) and the combination of the ADM with the CTM, referred to as SpecAugment [11] to increase the variability of laryngeal speech characteristics.

The CTM is conducted by obscuring a sequence of consecutive frames ranging from t_0 to $t_0 + t$, where t is randomly selected from the uniform distribution between zero to the designated time mask value T. In the ADM, the mel-frequency channels between f_0 to $f_0 + f$ are concealed, where f is randomly selected from the uniform distribution between zero to the designated frequency mask value F. Finally, the SpecAugment applies both previously described time and frequency masks to regular speeches. Let m_t and m_f be the CTM and ADM,respectively. We separately apply this mask to the input regular speech as

$$x_m = x \cdot m, \tag{2}$$

where m can be either m_t, m_f, or m_s and $\cdot$ is an element-wise product.

We concatenate x_m with m in a channel-wise manner before feeding it to the MaskCycleGAN-VC. The generator tries to fill in the masking frame to obtain its first synthesized alaryngeal speech, which is

$$y_s = G(concat(x_m, m)), \tag{3}$$

where y_s is the synthesized alaryngeal speech, $G(\cdot)$ is a voice converter function of the MaskCycleGAN-VC. Next, the cyclic conversion approach is activated by regenerating x from y_s, which is

$$x_r = G(concat(y_s, m_I)), \tag{4}$$

where x_r is the regenerated mel-spectrogram and m_I is the all-ones mask meaning no masking effect applying at this stage. Then, we compute

$$L_{cyc}^{x \longrightarrow y_s \longrightarrow x} = E_{x,m}(|x_r - x|), \tag{5}$$

where $L_{cyc}^{x \longrightarrow y \longrightarrow x}$ is a cycle-consistency loss for a cyclic conversion from x to itself and $E_{x,m}$ is the expectation over random mask selection and different inputs.

We repeat the above process but the input for the conversion will be the original alaryngeal speech y. This leads to another cycle-consistency loss

$$L_{cyc}^{y \longrightarrow x_s \longrightarrow y} = E_{y,m}(|y_r - y|), \tag{6}$$

where $L_{cyc}^{y \longrightarrow x_s \longrightarrow y}$ is a cycle-consistency loss for a cyclic conversion from y to itself,x_s is the synthesized normal speech, and $E_{y,m}$ is the expectation over random mask selection and different inputs. We use $L_{cyc}^{x \longrightarrow y_s \longrightarrow x}$ and $L_{cyc}^{y \longrightarrow x_s \longrightarrow y}$ in Eq.(9). Figure 1, 2 and 3 show the MaskCycleGAN-VC with CTM, ADM, and SpecAugment, respectively.

3 TIMBRE-ENHANCED LOSS

We introduce a timbre similarity metric to the MaskCycleGAN-VC cost function to obtain more authentic synthesized alaryngeal speeches. Timbre similarity refers to the perceived similarity in the quality of sounds, even when they have different pitches or durations. Timbre is one of the essential attributes of sound, and it allows us to distinguish between different voices even when they are playing the same note at the same volume. We first extract

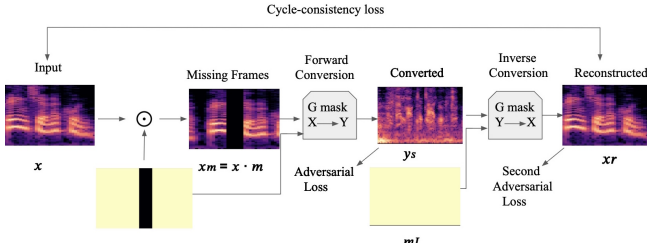

Figure 1: Pipeline of FIF using CTM masking method.

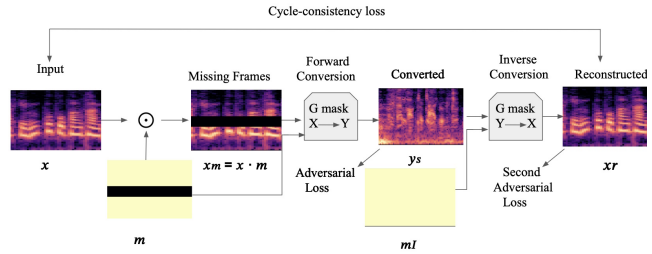

Figure 2: Pipeline of FIF using ADM masking method.

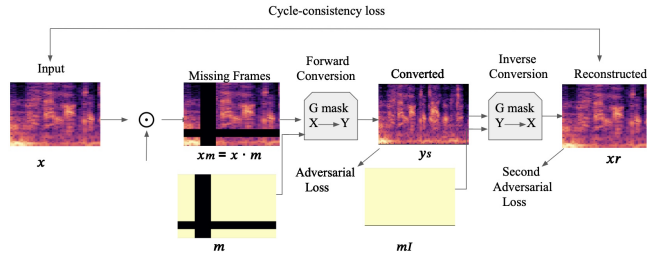

Figure 3: Pipeline of FIF using SpecAugment masking method.

speaker embeddings from synthesized alaryngeal and actual alaryngeal speeches using Resemblyzer[12] to calculate timbre similarity. It produces a summary vector of 256 values (embeddings) summarizing the voice's characteristics. Let v_s and v_a be the embedding vectors of synthesized and actual alaryngeal speeches, respectively. Next, we compute the cosine similarity between these two vectors via

$$T_{sim}(v_s, v_a) = \frac{v_s \cdot v_a}{|v_s||v_a|}, \tag{7}$$

where $T_{sim}(v_s, v_a)$ is the timbre similarity between v_s and v_a. Next define the timbre-enhanced loss as

$$L_{te} = 1 - T_{sim}(v_s, v_a). \tag{8}$$

We integrate this loss function and its weighting parameter λ_{te} to the total loss function of the MaskCycleGAN-VC as

$$L_{total} = L_adv + \lambda_{cyc} L_{cyc} + \lambda_{id} L_{id} + L_{adv2} + \lambda_{te} L_{te}, \tag{9}$$

4 EXPERIMENTAL RESULTS

4.1 Dataset

We conducted voice conversion experiments using a dataset comprising our recordings of Thai alaryngeal and normal speeches. The dataset comprises ten speakers with Thai alaryngeal speeches and 563 audio files of normal speeches. We utilized a subset of 1250 audio files from the alaryngeal speech speakers and 563 audio files from the normal speech speakers for training the MaskCycleGAN-VC. To evaluate the performance, we reserved 125 audio files from the alaryngeal speech speakers and another set of 125 audio files from the normal speakers. These evaluation sets were entirely distinct from the training data, ensuring our model's performance was assessed on unseen samples.

4.2 Training Settings

The architecture of MaskCycleGAN-VC used in this paper is the same as that in [4]. To be concise,the architecture of the voice converter is a 2-1-2D CNN [4] and the discriminator is PatchGAN [9]. Most training settings follows those from [4], which will be described as follows. Mel-spectrograms underwent normalization using training dataset statistics. Employing a least-squares GAN [10] as our GAN objective, the training process spanned 5000 iterations, with an Adam optimizer employed. Specifically, the converter and discriminator utilized learning rates of 0.0002 and 0.0001, respectively. Momentum terms were established as β_1 = 0.5 and β_2 = 0.999. The batch size was defined as 1, where individual training instances encompassed 64 randomly cropped frames, approximately equating to a duration of 0.75 seconds. Hyperparameters, λ_{cyc}, λ_{id}, and λ_{te} were configured at values of 10, 5, and 0.001,respectively.

As mentioned in Section 3, this paper includes the timbre-enhanced loss in the overall loss optimization during the training of Masked CycleGAN-VC to increase the similarity of speech characteristics between generated alaryngeal speeches and the actual ones. After incorporating Timbre Similarity to the MaskCycleGAN-VC cost function, the timbre similarity of all masking methods increased as displayed in Table 1.

Table 1: Timbre Similarity Evaluation with and without Timbre Enhanced Loss (L_{te}) for Different Masking Methods

Masking Methods	Timbre Similarity	
	Without L_{te}	With L_{te}
CTM	0.733	0.782
ADM	0.667	0.730
SpecAugment	0.729	0.750

4.3 Objective Evaluation

We evaluated the synthesized alaryngeal speeches using various metrics including Mel Cepstral Distorition (MCD), Frechet Audio Distance (FAD), Timbre Similarity and F0 comparison.

4.3.1 Mel Cepstral Distorition. The significance of MCD [8] as a metric in voice conversion from normal to alaryngeal lies in its role as a quantitative MCD measure for assessing the extent of

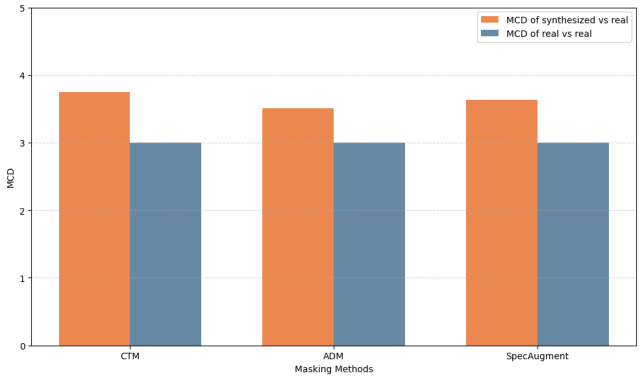

Figure 4: MCD scores between the actual and synthesized alaryngeal speeches under different masking methods.

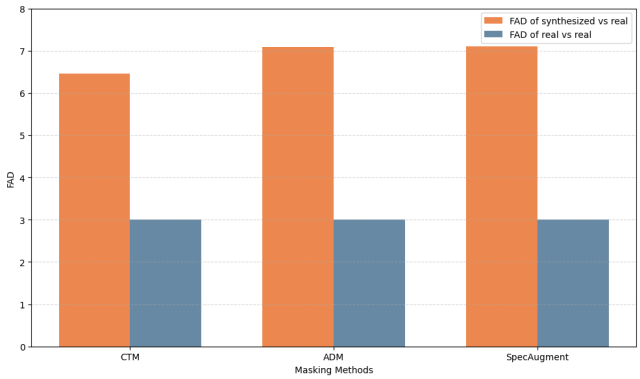

Figure 5: FAD scores between the actual and synthesized alaryngeal speeches under different masking methods.

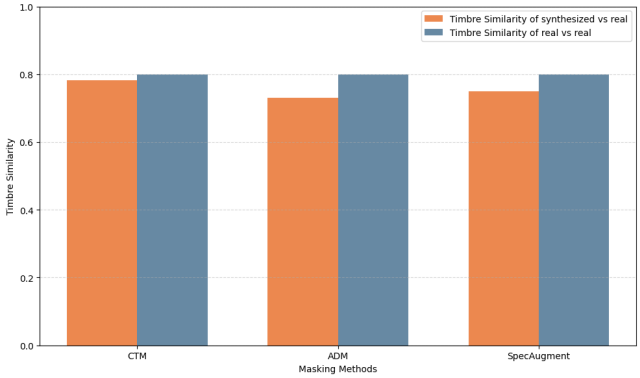

Figure 6: Timbre Similarity Scores between the actual and synthesized alaryngeal speeches under different masking methods.

spectral and acoustic differences between the original and synthesized alaryngeal speech. Suppose that y and y_s are the actual and synthesized mel-spectorgram of alaryngeal speeches, respectively. Then, the MCD can be defined as

$$\text{MCD}(y, y_s) = \frac{10}{ln(10)} \sqrt{2 \sum_{t=1}^{T} ||y(t) - y_s(t)||}, \tag{10}$$

where T is a number of frames in y, and $y(t)$ and $y_s(t)$ are the mel-spectral of the actual and the synthesized alaryngeal speeches at time t.

Figure 4 depicts the MCD scores of three masking methods. Upon our observation, the MCD values of the generated and the actual ones are in the same ranges of 3-4. Furthermore, we calculate the MCD scores between separate alaryngeal speakers to verify the congruence of their dissimilarity scores with our computed value. The acquired scores also range around three. The results indicate that our synthesized alaryngeal speeches are similar to the actual ones.

4.3.2 FAD. Frechet Audio Distance (FAD) [7] measures the dissimilarity between the spectral characteristics of the actual and synthesized alaryngeal speeches. Specifically, it deploys the mean and covariance of the extracted features obtained from sets of actual and synthesized alaryngeal speeches. The FAD can be defined as

$$F(y, y_s) = \|\mu - \mu_s\|^2 + tr(Cov + Cov_s - 2\sqrt{CovCov_s}), \tag{11}$$

where μ and Cov are a mean vector and a covariance matrix obtained from extracted features of a set of the actual alaryngeal speeches. μ_s and Cov_s are a mean vector and a covariance matrix obtained from extracted features of a set of the synthesized alaryngeal speeches. The obtained FAD values are illustrated in Fig. 5. We can observe that the FAD values between actual and synthesized alaryngeal speech sets are in the acceptable range. When we compare them with the FAD values from different actual alaryngeal speakers, We notice that the FAD values are below ten, which indicates that our alaryngeal synthesizer can produce realistic alaryngeal speeches.

4.3.3 Timbre Similarity. Following the timbre enhanced loss's incorporation into our loss function, we compared the timbre similarity scores between the converted and real alaryngeal speech as shown in Figure 6. Our method can give the similarity scores close to the genuine alaryngeal speakers ones.

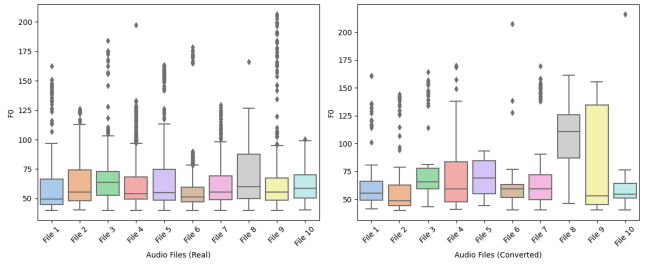

Figure 7: Distribution of F0 for Real and Converted Alaryngeal Speech.

4.3.4 Fundamental Frequency Comparison. We carried out an analysis of the fundamental frequencies in a set of ten audio files, encompassing both the authentic reference speeches and the artificially generated speeches. Figure 7 visually represents the fundamental frequency (F0) across both the genuine reference speech and the generated speech.

Specifically, we observed that the minimum frequency in both generated and actual alaryngeal speech is around 40 while the highest fundamental frequency in both speeches is around 200.

5 CONCLUSION

In conclusion, our paper has tackled the challenge of data augmentation in alaryngeal speech processing. We achieved this by employing innovative masking techniques like Articulatory Dimension Masking (ADM) and the novel SpecAugment method, leading to significant improvements in voice conversion quality and performance within the MaskCycleGAN-VC framework. Our approach not only enhances voice characteristic variability within the converted alaryngeal speech dataset but also introduces a groundbreaking concept: the Timbre Enhanced Loss. This dynamic score guides voice transformation to preserve crucial timbral characteristics. Our results underscore the effectiveness of our data augmentation technique, offering promising potential for boosting deep learning models in alaryngeal speech recognition and synthesis

REFERENCES

[1] Dario Amodei et al. 2016. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. 173–182.

[2] Mikołaj Bińkowski, Jeff Donahue, Aidan adn Elsen Erich Dieleman, Sander adn Clark, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. 2020. High Fidelity Speech Synthesis with Adversarial Networks. (2020). https://openreview.net/forum?id=r1gfQgSFDr

[3] Beiming Cao, Kristin Teplansky, Nordine Sebkhi, Arpan Bhavsar, Omer T. Inan, Robin Samlan, Ted Mau, and Jun Wang. 2022. Data augmentation for end-to-end silent speech recognition for laryngectomees. In *Proceedings of the Interspeech*.

[4] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion. In *Proc. ICASSP*. 6820–6824.

[5] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2020. Cyclegan-VC3: Examining and improving Cyclegan-VCs for Mel-Spectrogram Conversion. In *Interspeech 2020*.

[6] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2021. Maskcyclegan-vc: Learning non-parallel voice conversion with filling in frames. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[7] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*.

[8] Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Vol. 1.

[9] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with Markovian generative adversarial networks. In *Proc. ECCV*. 702–716.

[10] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proc. ICCV*. 2794–2802.

[11] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[12] Resemble-AI. 2021. Resemblyzer: A Python package to analyze and compare voices with deep learning. https://github.com/resemble-ai/Resemblyzer. Accessed: Aug. 24, 2023.

[13] Marinela Rosso, Ljiljana Sirić, Robert Tićac, Radan Starcević, Igor Segec, and Nikola Kraljik. 2012. Perceptual evaluation of alaryngeal speech. *Collegium Antropologicum* 36, Suppl 2 (2012), 115–118.