Few-shot Speaker Adaptation for Text-to-Speech Synthesis Using Non-Target Speaker Corpora for Glossectomy Patients

Masayori Okamura*, Masanobu Abe* and Sunao Hara*
* Okayama University, Japan
E-mail: m_oka@a.cs.okayama-u.ac.jp, {abe-m, hara}@okayama-u.ac.jp

***Abstract*—For individuals with articulation disorders, such as patients after subtotal glossectomy, text-to-speech (TTS) provides a means of voice communication. To preserve personal identity in TTS, speaker adaptation methods are employed. However, when the target's disordered speech is used for adaptation, there is a trade-off: while the speaker's identity can be reproduced, speech intelligibility tends to deteriorate. To address this challenge, we propose a speaker adaptation method that uses less than one minute of the target speaker's disordered speech. The proposed method performs adaptation using not only the target's disordered speech but a parallel corpus of healthy and disordered speech from non-target speakers. This approach enables the model to learn speaker features that are robust to the intelligibility of the reference speech. Evaluation results show that the proposed method significantly improves the character accuracy of synthesized speech compared to the baseline, achieving intelligibility levels comparable to healthy speech. Furthermore, subjective evaluations demonstrate an improvement in the mean opinion score (MOS) for speaker similarity. Visualization of speaker embeddings confirmed that our method brings the distributions of a single speaker's healthy and disordered speech closer together. These findings demonstrate that the proposed method is effective for achieving high-quality, personalized speech synthesis from limited and disordered speech samples[1].**

I. Introduction

Vocal communication plays a vital role in human social interaction. However, individuals with articulation disorders, particularly those who have undergone subtotal glossectomy, face significant challenges in speaking. Subtotal glossectomy involves the removal of more than half of the tongue, typically to treat conditions such as tongue cancer. As a result, affected individuals experience difficulty articulating fricatives and plosives due to impaired tongue function[1]. The development of alternative communication tools for these patients that can substitute for natural speech is of great importance.

Speech synthesis technologies such as text-to-speech (TTS) and voice conversion (VC) offer an alternative means of vocal communication for glossectomy patients. Previous studies have proposed both VC and TTS-based approaches for this population.

In VC approaches[2][3], models are trained to convert disordered speech into healthy-sounding speech. These VC models can enhance intelligibility of the target speaker's speech without requiring manual input devices, thereby supporting real-time communication. However, training the VC model demands a substantial amount of speech data from the patient, imposing a physical burden. Moreover, conventional VC approaches often require parallel corpora comprising both healthy and disordered speech from the same target speaker, a practical limitation.

In contrast, TTS approaches[4][5] involve fine-tuning a multi-speaker TTS model using the target speaker's disordered speech. TTS systems benefit from large pre-trained models based on healthy speech and can incorporate explicit textual instructions, making it easier to improve intelligibility than VC methods. However, these approaches still face a trade-off between speaker similarity and intelligibility: greater adaptation to the target voice degrades intelligibility. Furthermore, improving speaker similarity typically requires a large amount of target speaker data, similar to VC systems.

Recently, attention has shifted toward zero-shot and few-shot TTS, with the aim of synthesizing speech with controllable speaker characteristics using minimal data from the target speaker. To control speaker characteristics in these approaches, typically one of two strategies is employed: (1) using intermediate layer outputs from speaker verification models[6], or (2) jointly training a speaker encoder with the TTS model to generate speaker embeddings[7][8]. The first strategy benefits from robustness to unseen speakers due to previous training on a wide range of speaker identities. The second can produce embeddings optimized for the TTS task, capturing features more suitable for speech synthesis than those extracted via speaker verification models.

In this study, we propose a few-shot speaker adaptation method for synthesizing speech in glossectomy patients, using a retrainable speaker encoder. Our preliminary investigation indicates that conventional speaker embeddings often reflect clear distinctions between disordered and healthy speech from the same speaker. To overcome this, we introduce a training strategy that aligns disordered speech embeddings with those of healthy speech. During speaker adaptation, we augment the disordered speech of the target speaker with speech from other speakers. In addition, to explicitly guide synthesis toward healthy speech at inference time, we employ different phoneme labels when training on healthy and disordered speech. Our

[1]Audio samples are available at https://yori2000.github.io/APSIPA-demo/

979-8-3315-7206-8/25/$31.00

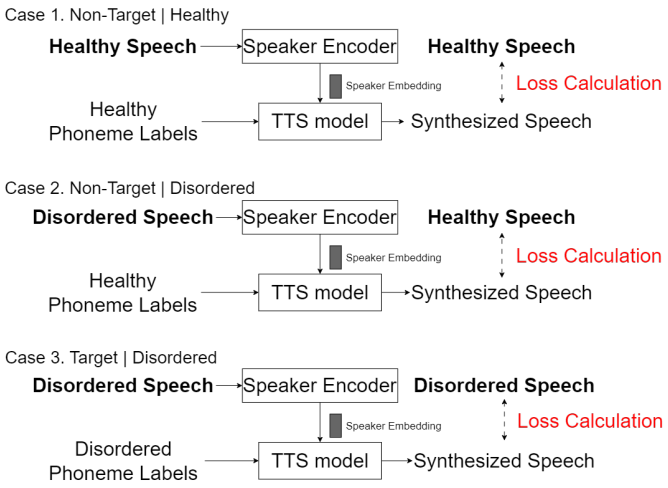


Fig. 1: Flowchart of input to loss computation for each case of reference speech during speaker adaptation.

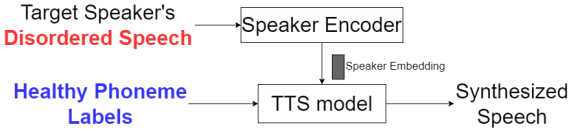


Fig. 2: Inference stage

experimental results show that the proposed method improves both speaker similarity and intelligibility, compared to adaptation based only on disordered speech. We also demonstrate that the method brings the distributions of a single speaker's healthy and disordered speech embeddings closer together.

II. Proposed Method

A. Conditional Phoneme Labeling for Feature Preservation

The phoneme-prosody labels used in the proposed method comprise a shared set of prosodic labels and two distinct sets of phoneme labels: one for healthy speech and one for disordered speech. Table I provides examples of the phoneme-prosody labels employed in the proposed method. We assume that the speech of healthy speakers and that of glossectomy patients exhibit different acoustic characteristics, even when producing the same phoneme. Accordingly, to preserve healthy phonetic features during speaker adaptation, we apply condition-specific phoneme labels as described in Section II-B. This distinction is applied across all phonemes rather than only to specific consonants or vowels. This approach is necessary due to the often imprecise phoneme alignments in disordered speech, which can affect adjacent phonetic features. This label distinction strategy prevents the TTS model from confusing the acoustic characteristics of the healthy and disordered speech, allowing it to retain the phonetic features of healthy speech throughout the speaker adaptation process.

B. Training and Inference

The training process consists of two stages: pre-training and speaker adaptation. Table II summarizes the input and ground-truth data used during training and inference in the proposed TTS model. The goal of pre-training is to equip the TTS model with a general-purpose ability to synthesize intelligible speech. To this end, we used a multispeaker speech corpus comprising healthy speech for pre-training. The input consisted of healthy-phoneme labels and corresponding healthy speech.

The goal of speaker adaptation is twofold: (1) to maintain the intelligible speech synthesis capability acquired during pre-training, and (2) to learn speaker characteristics from the target speaker's disordered speech. To achieve this, we use a dataset that combines the target speaker's disordered speech with data from non-target speakers. The non-target speaker data used for speaker adaptation consists of a parallel corpus of healthy and disordered speech. This dataset is a subset of our four-speaker, in-house corpus (described in Section III-C) , and includes the data from the three speakers excluding the target speaker.

Figure 1 illustrates the training flow for speaker adaptation. There are three training cases for the TTS model:

- Case 1: When healthy speech from non-target speakers is used as the reference, the input includes healthy-phoneme labels and the reference speech. The same reference speech is also used as the ground-truth for loss computation.
- Case 2: When disordered speech from non-target speakers is used as the reference, the input still comprises healthy-phoneme labels and disordered reference speech. The corresponding parallel healthy speech is used as the ground-truth.
- Case 3: When disordered speech from the target speaker is used as the reference, both the input and ground truth consist of disordered speech, accompanied by disordered-phoneme labels.

In Cases 1 and 2, healthy speech is used as the ground truth, regardless of the reference condition. Case 2 directly trains the model for the inference-time task of synthesizing healthy speech from disordered reference inputs. Simultaneous training with Case 1 helps the model learn ideal speaker embedding from healthy speech, which serves as the target representation for Case 2. This process enables the speaker encoder to disentangle speaker identity from intelligibility, thereby learning an intelligibility-invariant speaker representation.

Figure 2 shows the inference stage. Here, the model receives the target speaker's disordered speech as the reference and healthy-phoneme labels as the text input. Although the model has not encountered each input combination, it has been trained on both components independently. Therefore, this inference setting encourages the model to synthesize intelligible speech with the target's speaker characteristics.

III. Experiments

A. Architecture

The proposed method is based on VITS[9], an end-to-end TTS model, with the addition of a speaker encoder. Specifically, we adopt the speaker encoder introduced by Fujita et al.[8], which is based on a self-supervised learning (SSL) model. The speaker encoder architecture is shown in Figure

TABLE I: Phoneme-prosody labels.

Text	イランに天気予報はない。
Healthy-phoneme label	^ i] r a N n i _ t e [N k i y o] h o o w a n a i $
Disordered-phoneme label	^ i'] r' a' N' n' i' _ t' e' [N' k' i' y' o'] h' o' o' w' a' n' a' i' $
Text	変わりびなは世相をつづる。
Healthy-phoneme label	^ k a [w a r i b i] n a w a # s e [s o o o # ts u [z u r u $
Disordered-phoneme label	^ k' a' [w' a' r' i' b' i'] n' a' w' a' # s' e' [s' o' o' o' # t' s' u' [z' u' r' u' $

TABLE II: Input data and ground-truth data for training and inference of the TTS model.

	Pre-Training	Speaker Adaptation		Inference
		Non-Target Speaker	Target Speaker	
Reference Speech	Healthy	Healthy / Disordered	Disordered	Disordered
Phoneme Label	Healthy	Healthy	Disordered	Healthy
Ground-Truth Speech	Healthy	Healthy	Disordered	-

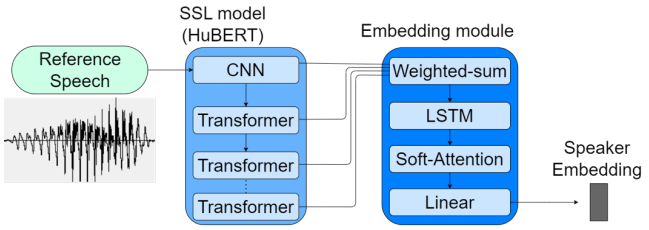


Fig. 3: Speaker encoder architecture.

TABLE III: Hyperparameters.

	Pre-training	Speaker adaptation
Optimization algorithm	AdamW	AdamW
Learning rate	1.0×10^{-4}	1.0×10^{-5}
Batch size	12	12
Epochs	1000	500

TABLE IV: Datasets

	JSUT	JVS	In-house Corpus
Speech Type	Healthy	Healthy	Healthy / Disordered
Number of Utterances	7,306	13,000	212
Total Duration	10 hours	26 hours	16 minutes
Speakers	1 Female	49 Males, 51 Females	3 Males, 1 Female
Sampling Frequency	48 kHz	24 kHz	20 kHz

3. For the SSL model, we use the pre-trained HuBERT-base[2][10][11] provided by Rinna Inc. During TTS training, the HuBERT weights are kept fixed, only the subsequent layers are fine-tuned. The speaker encoder outputs a speaker embedding that conditions the TTS model, allowing it to control speaker characteristics. Table III lists the hyperparameters used in both the pre-training and speaker adaptation phases.

B. Datasets

Table IV summarizes the datasets used in the experiments. For pre-training of the TTS model, we employed the JSUT[12] and JVS[13] corpora, both of which comprise healthy speech. For speaker adaptation, we used the in-house corpus (detailed in Section III-C) as follows: 10 pseudo-disordered speech utterances from the target speaker, and healthy / pseudo-disordered speech from the three remaining speakers, based on the speaker adaptation conditions described in Section III-D. All audio was resampled to 22050 Hz for the input into the VITS model. For the speaker encoder, audio was resampled to 16000 Hz to match the pre-trained HuBERT sampling frequency. Text for each dataset was processed using pyopenjtalk.

[2]https://huggingface.co/rinna/japanese-hubert-base

C. Pseudo-disordered Speech

In speech synthesis evaluation, it is ideal to have access to the target speaker's healthy speech as ground truth. However, when the target speaker is a glossectomy patient, such speech data is typically unavailable. Furthermore, from the perspective of physical strain, it is challenging for patients to provide extensive recordings.

To address this, we use an in-house corpus from a previous study[14], which includes both healthy and pseudo-disordered speech. In this corpus, pseudo-disordered speech was recorded by healthy speakers using a device that restricts tongue movement to simulate the articulation patterns of glossectomy patients. In our experiment, one of the four speakers is designated as the target speaker, for whom only the pseudo-disordered speech is used. The remaining three speakers contribute both healthy and pseudo-disordered speech for the purpose of speaker adaptation.

D. Training Conditions for Speaker Adaptation

During the speaker adaptation phase, each training set was built by supplementing a baseline of 10 disordered utterances from the target speaker with additional data, as outlined in Table V.

- **C1** *(Baseline)*: Includes only 10 disordered utterances from the target speaker.
- **C2** Expands C1 by adding one healthy utterance from each of 100 speakers in the JVS corpus, using parallel sentence (file ID: 'voiceactress100_001').
- **C3** Builds on C1 using a non-parallel set of 100 healthy utterances from the JVS corpus, corresponding to 'voiceactress100_001' through 'voiceactress100_100'.
- **C4** Supplements C1 with both healthy and disordered speech from three non-target speakers in the in-house corpus. Here, both types of speech are used as reference inputs, only healthy speech is used as the ground truth.
- **C5** Expands on C4 by adding the same parallel subset from JVS as in C2.
- **C6** Expands on C4 by including the non-parallel JVS subset from C3.

TABLE V: Data for speaker adaptation

	JVS	In-house Corpus
C1: Disordered only	-	-
C2: C1 + JVS(Parallel)	100 spk, 1 utt/spk (Parallel text)	-
C3: C1 + JVS(Non-parallel)	100 spk, 1 utt/spk (Non-parallel text)	-
C4: C1 + Non-target speech	-	50 disordered + 50 healthy utt. from 3 spk
C5: C4 + JVS(Parallel)	100 spk, 1 utt/spk (Parallel text)	50 disordered + 50 healthy utt. from 3 spk
C6: C4 + JVS(Non-parallel)	100 spk, 1 utt/spk (Non-parallel text)	50 disordered + 50 healthy utt. from 3 spk

E. Evaluation Experiment

To evaluate the effectiveness of the proposed method in terms of intelligibility, we used the character accuracy of synthesized speech as the objective metric. For each experimental condition described in Section III-D, speech was synthesized using the corresponding speaker-adapted TTS model. The synthesized speech was then input into an ASR system to compute character accuracy. The ASR system used was OpenAI's Whisper large-v3[3]. Character accuracy was calculated from the character error rate (CER) using the following formula:

$$\text{Accuracy} = 1 - \text{CER} = 1 - \frac{S + D + I}{N} \tag{1}$$

where S, D, and I denote the number of substitutions, deletions, and insertion errors, and N is the total number of characters in the ground-truth transcript. A value closer to 1.0 indicates higher intelligibility of the synthesized speech.

In addition to the objective evaluation, we conducted a subjective evaluation of speaker similarity using a Mean Opinion Score (MOS) study. Participants included 11 native speakers (9 male, 2 female), all in their 20s. In each trial, subjects were asked to rate the speaker similarity between synthesized speech and a reference utterance from the target speaker on a 5-point scale. The text content of the synthesized and reference speech always differed. The scale, adapted from Jia et al.[15], was defined as follows: 1 (Not at all similar), 2 (Slightly similar), 3 (Moderately similar), 4 (Very similar), and 5 (Extremely similar). The reference speech consisted of healthy utterances from the four speakers in the in-house corpus (see Table IV). While the synthesized speech was generated by TTS models adapted under each condition (see Table V). We selected six sentences and formed three evaluation pairs per speaker, with different textual content. The order of text content and reference / synthesized speech presentation was randomized. Each condition involved a total of 264 evaluations: (3 pairs $\times$ 2 substitutions $\times$ 4 target speakers $\times$ 11 subjects).

IV. Results

A. Objective Evaluation of Speech Intelligibility

Table VI presents the results of the objective evaluation of speech intelligibility. In addition to the scores for each experimental condition (C1–C6), character accuracy results are also shown for the original healthy and disordered speech of the four in-house speakers, providing reference points.

Across all experimental conditions, a notable improvement in character accuracy was observed, compared to the original disordered speech. This improvement can be attributed to two key factors: the pre-training of the TTS model on a sufficient amount of healthy speech to establish baseline level of intelligibility, and the use of separate input labels for disordered and healthy speech.

To examine the importance of utterance diversity in the supplementary data, we first compare conditions C1, C2, and C3. While C1 outperformed Original_Disordered, its accuracy remains considerably lower than that of the Original_Healthy across all speakers. This indicates that training solely on disordered speech leads to catastrophic forgetting, where the TTS model gradually loses its capacity to synthesize intelligible speech. The comparison between C1 and C2 revealed no significant difference in accuracy, implying that merely increasing the volume of healthy speech or the number of speakers offers limited benefits unless accompanied by linguistic diversity. In contrast, C3 significantly outperformed C2, indicating that linguistic diversity, rather than the parallel nature of the data, is essential for maintaining intelligible synthesis. These findings highlight that diverse utterance content allows the model to better preserve the acoustic features of each healthy phonemes.

To assess the effect of the in-house corpus, we turn to conditions C4, C5, and C6. C4 shows markedly better accuracy than C1, demonstrating that including utterance diversity from the in-house corpus improves intelligibility. However, adding JVS data(as in C5 and C6) does not yield further gains over C4. This implies that the utterance diversity already present in the in-house non-target data is sufficient, and that further augmentation with JVS does not significantly enhance performance in terms of intelligibility.

B. Subjective Evaluation of Speaker Similarity

Table VII shows the results of the subjective evaluation of speaker similarity. Based on the findings of the objective evaluation, we selected conditions C1 and C4 for this experiment. As an additional baseline, we included a zero-shot model, in which the pre-trained TTS model is used directly without any speaker adaptation.

As expected, the zero-shot model received the lowest average score (1.62). Condition C1, which was adapted using only the target speaker's disordered speech, achieved a moderate score of 3.27. However, this result is not particularly convincing given the poor intelligibility performance under C1 reported in Table VI. In contrast, the proposed method (C4) achieved a higher score of 3.89, clearly outperforming both zero-shot and C1. These results confirm that the proposed speaker adaptation approach is more effective at capturing speaker identity, even when disordered speech is used as refer-

[3]https://github.com/openai/whisper

TABLE VI: Character accuracy of synthesized speech for each speaker(MHM, MKO, MKT, FAH) under each experimental condition(C1–C6). Results are presented as mean ± 95% confidence interval.

	MHM	MKO	MKT	FAH	Total
Original_Disordered	0.22 ± 0.07	0.15 ± 0.07	0.36 ± 0.07	0.35 ± 0.07	0.27 ± 0.04
Original_Healthy	0.92 ± 0.03	0.91 ± 0.03	0.91 ± 0.03	0.91 ± 0.03	0.91 ± 0.01
C1: Disordered only	0.45 ± 0.09	0.41 ± 0.09	0.53 ± 0.09	0.50 ± 0.09	0.47 ± 0.04
C2: C1 + JVS (parallel)	0.50 ± 0.08	0.40 ± 0.09	0.49 ± 0.07	0.54 ± 0.08	0.48 ± 0.04
C3: C1 + JVS (non-parallel)	0.84 ± 0.04	0.87 ± 0.04	0.87 ± 0.04	0.87 ± 0.04	0.86 ± 0.02
C4: C1 + Non-target speech	0.91 ± 0.03	0.91 ± 0.03	0.90 ± 0.03	0.87 ± 0.04	0.90 ± 0.02
C5: C4 + JVS (parallel)	0.89 ± 0.03	0.89 ± 0.04	0.90 ± 0.03	0.89 ± 0.03	0.89 ± 0.02
C6: C4 + JVS (non-parallel)	0.90 ± 0.03	0.90 ± 0.04	0.90 ± 0.03	0.89 ± 0.03	0.90 ± 0.02

TABLE VII: Mean Opinion Score (MOS) for speaker similarity. Results are shown as mean ± 95% confidence interval.

	MHM	MKO	MKT	FAH	Total
Ground Truth	4.91 ± 0.09	4.79 ± 0.17	4.98 ± 0.03	4.92 ± 0.09	4.90 ± 0.05
zero-shot	1.73 ± 0.36	1.50 ± 0.32	1.98 ± 0.47	1.27 ± 0.28	1.62 ± 0.18
C1: Disordered only	3.13 ± 0.46	2.70 ± 0.50	2.83 ± 0.39	4.42 ± 0.30	3.27 ± 0.28
C4: C1 + Non-target speech	4.41 ± 0.27	3.79 ± 0.37	2.83 ± 0.56	4.55 ± 0.32	3.89 ± 0.27

ence. In conclusion, the results of both objective and subjective evaluations demonstrate that our method successfully enhances speaker similarity while maintaining high speech intelligibility.

C. Discussion

To investigate the factors contributing to the improvement in speaker similarity achieved by the proposed method, we visualized and compared the distributions of speaker embeddings across diffent model configurations. While speaker encoder cosine similarity (SECS)[6] is commonly used as a speaker similarity metric, external speaker encoders like ECAPA-TDNN are susceptible to variations in speech intelligibility as mentioned in Section I. Therefore, instead of using SECS, we visualized the speaker embeddings from our adapted models. We analyzed embeddings from the following four sources:

- The output of an intermediate layer from a conventional speaker identification model (ECAPA-TDNN).
- The speaker encoder adapted under condition C1.
- The speaker encoder adapted under condition C4.
- The speaker encoder adapted under condition C7.

Condition C7 serves as an ablation model adapted using the same dataset as C4, but only with training Cases 2 and 3 (shown in Figure 1), omitting Case 1. In other words, the C7 model was trained using only the inference time configuration, in which disordered speech was used as reference. Therefore, comparing C4 and C7 isolates the effect of including Case 1 in our full speaker adaptation strategy.

Figure 4 shows the distributions of speaker embeddings from C1, C4, and C7. For comparison, Figure 4a shows the speaker embedding produced by ECAPA-TDNN[4][17], a conventional speaker embedding model. Each embedding was obtained by feeding both healthy and disordered speech from each speaker into the respective encoders. The proximity of a speaker's healthy and disordered embeddings reflects the robustness of the speaker's representation of intelligibility variations.

Compared to both ECAPA-TDNN and C1, the speaker embedding distribution for C4 exhibits substantially tighter clustering between a speaker's healthy and disordered speech. This indicates that our method enables the speaker encoder to extract speaker identity robustly, even when disordered speech is used at reference.

Furthermore, C4 shows tighter clustering than C7, suggesting that the inclusion of Case 1 during adaptation is crucial to improving speaker similarity. This supports the effectiveness of the proposed full adaptation strategy in learning intelligibility-invariant speaker embeddings.

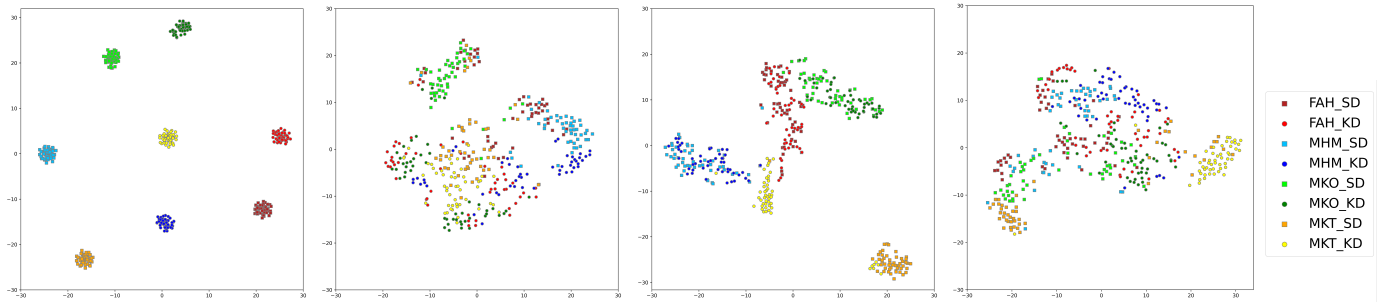
However, the speech of MKT could not achieve improved in the clustering. We speculate that the acoustic differences between the MKT's healthy and disordered speech are distinct from those of the other speakers, therefore, the modeling using the parallel corpus from the non-target speakers was not sufficiently functioned for MKT.

V. Conclusion and Future Work

In this study, we propose a speaker adaptation method for TTS tailored to glossectomy patients, requiring less than 1 minute of the target speaker's speech. To enable effective speaker adaptation from disordered speech while preserving intelligibility, we introduced two key techniques. The first distinguishes between phoneme labels of healthy and disordered speech, preserving clear phonetic features during training. The second involves an adaptation strategy that leverages both healthy and pseudo-disordered speech from non-target speakers to robustly extract speaker characteristics from disordered input. Evaluation experiments confirmed the effectiveness of the proposed approach. In the speech intelligibility evaluation, the proposed condition (C4) significantly improved the character accuracy of synthesized speech, reaching levels comparable to original healthy speech. In the speaker similarity evaluation, the proposed method achieved higher scores than the baseline (C1), which was adapted only to the target's disordered speech. In addition, visualization of the speaker embedding showed that our method enabled the model to learn robust, intelligibility-invariant speaker representations.

In future work, a performance comparision with state-of-the-art zero-shot TTS models, such as YourTTS[7] and VALL-

[4]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb



(a) ECAPA-TDNN (b) C1 (c) C4 (d) C7

Fig. 4: Distribution of speaker embeddings for ECAPA-TDNN, C1, C4, and C7. The embedding vectors were obtained by inputting both healthy (SD) and disordered (KD) speech into the respective speaker encoder, and visualized after dimensionality reduction to two dimensions using the t-SNE algorithm[16]. In the legend, the first three characters represent the speaker ID.

E[18], is necessary. Furthermore, since the VITS architecture used in our proposed method can be easily adapted for VC tasks, we plan to explore a VC-based approach as well.

References

[1] A. Ohira, H. Yoshimasu, and T. Oyama, "Articulation dynamics after extensive glossectomy," *The Japanese Journal of Voice and Speech Disorders*, vol. 26, no. 3, pp. 215–223, 1985, (in Japanese).

[2] K. Tanaka, S. Hara, M. Abe, *et al.*, "Speaker dependent approach for enhancing a glossectomy patient's speech via gmm-based voice conversion," in *Proc. Interspeech*, 2017, pp. 3384–3388.

[3] H. Murakami, S. Hara, M. Abe, *et al.*, "Naturalness improvement algorithm for reconstructed glossectomy patient's speech using spectral differential modification in voice conversion," in *Proc. INTERSPEECH*, 2018, pp. 2464–2468.

[4] T. Yoshimoto, R. Takashima, C. Sasaki, *et al.*, "Improving speech intelligibility for people with spinal muscular atrophy using speaker adaptation of acoustic models," in *The 2021 Autumn Meeting of the Acoustic Society of Japan*, (in Japanese), 2021, pp. 1053–1056.

[5] T. Yoshimoto, K. Matsubara, R. Takashima, *et al.*, "Improving speech intelligibility for people with spinal muscular atrophy using multi-speaker tts," in *The 2022 Spring Meeting of the Acoustic Society of Japan*, (in Japanese), 2022, pp. 1045–1048.

[6] E. Cooper, C.-I. Lai, Y. Yasuda, *et al.*, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP*, 2020, pp. 6184–6188.

[7] E. Casanova, J. Weber, C. D. Shulby, *et al.*, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022, pp. 2709–2720.

[8] K. Fujita, T. Ashihara, H. Kanagawa, *et al.*, "Zero-shot text-to-speech synthesis conditioned using self-supervised speech representation model," in *ICASSPW*, 2023, pp. 1–5.

[9] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[10] K. Sawada, T. Zhao, M. Shing, *et al.*, "Release of pre-trained models for the Japanese language," in *Proc. LREC-COLING*, 2024, pp. 13 898–13 905.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] R. Sonobe, S. Takamichi, and H. Saruwatari, *JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis*, arXiv preprint arXiv:1711.00354, 2017.

[13] S. Takamichi, K. Mitsui, Y. Saito, *et al.*, *JVS corpus: Free Japanese multi-speaker voice corpus*, arXiv preprint arXiv:1908.06248, 2019.

[14] H. Murakami, S. Hara, and M. Abe, "DNN-based voice conversion with auxiliary phonemic information to improve intelligibility of glossectomy patients'speech," in *Proc. APSIPA-ASC*, 2019, pp. 138–142.

[15] Y. Jia, Y. Zhang, R. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[16] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[17] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[18] S. Chen, C. Wang, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 705–718, 2025.