Investigation of Text-to-Speech-based Synthetic Parallel Data for Sequence-to-Sequence Non-Parallel Voice Conversion

Ding Ma, Wen-Chin Huang and Tomoki Toda
Graduate School of Informatics, Nagoya University, Nagoya, Japan
E-mail: {ding.ma, wen.chinhuang}@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Sequence-to-sequence (seq2seq) voice conversion (VC) can achieve high-quality VC performance with a parallel dataset, but it is still challenging when the parallel dataset is not available. One way to overcome this issue is using synthetic parallel data (SPD) produced by text-to-speech (TTS) models trained with source and target speakers' voices in an available VC dataset. In this paper, we conduct a systematic study on the effects of SPD on seq2seq VC performance. Some factors, such as a comparison of SPD by source or target speaker, the effects of SPD in a semiparallel setting including a parallel subset, and the usage of SPD with external text data are investigated. The results show the effects of SPD depend on TTS performance and VC training datasize; i.e., 1) when the datasize is small, causing SPD quality degradation as the resulting TTS performance is limited, the training pair containing source SPD and target natural speech tends to yield slightly better VC results than that containing source natural speech and target SPD, 2) although SPD makes it possible to use a nonparallel dataset, using parallel subset is still effective, and 3) SPD with external text data as data augmentation can improve parallel seq2seq VC performance.**

I. Introduction

A speech mainly consists of two kinds of information: speaker identity and linguistic information. Voice conversion (VC) is a methodology that aims to convert the speaker identity of speech from source speaker into target speaker while preserving the linguistic information [1], [2], [3]. If VC can better capture and convert the features influencing speaker identity such as prosody, timbre, emotion and so on, the generated speech from the source speech will be more natural and closer to the target speech. However, conventional VC models follow frame-level mapping paradigm, which means the converted speech always gets the same temporal structure as that of the source speech, thus the loss of some important feature information during VC process will occur. Also, conventional VC systems hardly convert long-term dependencies well, such as prosody conversion including fundamental frequency (F0) patterns and phoneme duration, which limits the performance of VC.

In recent years, the sequence-to-sequence (seq2seq) models, which have emerged from the development of deep neural networks (DNN) [4], have been able to overcome the aforementioned shortcomings of conventional VC models. Modern seq2seq VC models are equipped with attention mechanism under an encoder-decoder framework [5], [6], [7], making the seq2seq VC models able to automatically determine the output phoneme duration based on what they have learned. Consequently, the seq2seq models are able to capture the long-term dependencies in VC, and they are better at converting the speaker identity than the frame-level mapping models [8].

In spite of the fact that the seq2seq VC models have the promising prospects, most of them rely on a large amount and parallel training corpus, which limits the further scope of application in practice. In order to make better use of the superiorities of seq2seq VC models, some attempts have been made to address this issue. A non-parallel seq2seq method was proposed in [9], which utilized a two-stage recognition encoder to extract and disentangle speaker and linguistic representations, whereas a complex rigorous preparation of hyperparameter tuning was needed. In [10], a framework by using connectionist temporal classification phonetic posteriorgrams (CTC-PPGs) to replace time-aligned PPGs and alleviate the impact of the prosody was proposed, but necessitated a high-precision CTC based automatic speech recognition model. In our previous study, a seq2seq VC model named Voice Transformer Network (VTN) based on the transformer architecture with TTS pre-training was proposed [11]. We extended it with synthetic parallel data (SPD), which was a more efficient approach compared to the way based on modifying the architecture of seq2seq model to tackle non-parallel data in the task of Voice Conversion Challenge 2020 (VCC2020) [12]. In VCC2020 task 1, the original dataset was semiparallel, contained both parallel and non-parallel subsets, and the proportion of parallel subset was very small. Hence, semiparallel can be regarded as the relaxation of non-parallel case. Although we have achieved relatively good VC results, there are still uncertainties about the usage of SPD on seq2seq VC model that need to be investigated.

In this paper, we conduct a systematic study focusing on the effects of SPD on seq2seq VC performance. We try to address the following questions:

Q1: What are the feasibility and properties of using SPD?

Q2: How can this method benefit from a semiparallel setting?

Q3: What are the influences of using external text data?

We carefully investigate several factors, such as a comparison of SPD by source or target speaker, the effects of SPD in a semiparallel setting, and the usage of SPD with external text data.

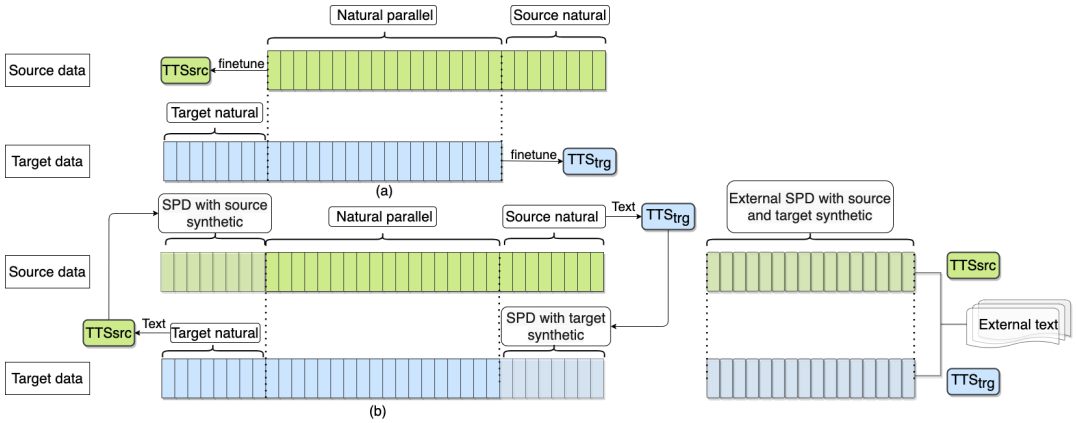

Fig. 1. Generation process of a synthetic parallel dataset (SPD) from a semiparallel dataset.
(a) TTS training process using the semiparallel dataset; and (b) SPD generation process using source synthetic data, target synthetic data, and external SPD.

The remaining part of the paper is structured as follows. The background and motivation are introduced in Section II. The experimental details and the discussion of the results are outlined in Section III. Conclusions from the study are presented in Section IV.

II. Background and Motivation

Based on the conditions of training data, VC can be classified into parallel and non-parallel ones [9], [13]. Parallel VC means that source and target speakers have same corpus content, which can provide more accurate alignment to help build a better mapping function of an acoustic model in a supervised manner. In contrast, non-parallel VC means that the corpus content is different, even being cross-lingual, and therefore, the mapping function of the acoustic model needs to be built in an unsupervised manner. Considering the difficulty to collect parallel data for VC, it is undoubtedly more practical to implement non-parallel VC, meanwhile it is challenging though.

Due to the semiparallel setting in VCC2020 task 1, it is necessary to use semiparallel dataset to provide as much data as possible for the VTN model. Thus, we use the SPD to convert the semiparallel issue into parallel one based on the work by [4], [14]. Here, the SPD is generated by TTS models. The generation process of SPD from the semiparallel dataset is shown in Fig. 1. First, a source speaker's TTS model and a target speaker's TTS model are separately developed by using their natural speech data in the semiparallel dataset. Finetuning approach is employed as the amount of available natural speech data is very limited. And then, source synthetic speech data and target synthetic speech data are generated by using the trained TTS models. Consequently, in addition to the original natural parallel data in the semiparallel set, two kinds of training pairs containing SPD are also generated, which are <source synthetic, target natural> and <target synthetic, source natural>. Moreover, it is possible to additionally introduce external text to produce external SPD. It is naturally motivated us to figure out a specific impact of using these three types of SPD on seq2seq VC training.

III. Experimental Evaluations for SPD Effect Investigation

The section is divided into five subsections. The first subsection introduces the overall experimental data configuration, and the next three subsections address the three questions Q1, Q2, and Q3 mentioned in Section I. Each subsection gives viewpoints and discussions based on results of objective evaluations. Finally, the last subsection presents discussions on these three questions based on results of subjective evaluations.

A. Datasets and configuration

In this study, the original datasets were from the CMU ARCTIC database [15], containing parallel 1132 utterances recorded by several English speakers in 16 kHz, where the female speakers of clb and slt, and male speakers of bdl and rms were selected as source or target speakers. Here, we used 100 utterances from the database as a development set and other 100 utterances as an evaluation set, respectively, and then, the remaining utterances were used as a training set. For the external data, we chose English corpus from M-AILABS database [16] with totally 15,369 utterances, roughly 30 hours long.

We followed the implementations of the open-source ESPnet toolkit [17], [18], where we used 80-dimensional mel filterbanks with 1024 FFT points and a 256 point frame shift to extract the acoustic features. The Transformer-TTS architecture [19] was used as a TTS model to generate SPD. Details of the model and training configuration can be found online[1]. As for the VC model, we directly used the VTN model and followed the official implementation[2]. These two models were both pretrained using M-AILABS database.

[1]https://gist.github.com/unilight/a48f99cf6a47c0b4e5b96fe1d6e59397
[2]https://github.com/espnet/espnet/tree/master/egs/arctic/vc1

TABLE I
THE COMPARISON RESULTS WITH DIFFERENT TRAINING PAIR AND DATASIZE. TTS-450, TTS-400, TTS-200 AND TTS-80 REPRESENT THE HOMOLOGOUS DATASIZE OF TTS FINETUNING, WHICH ALSO REFLECT TTS PERFORMANCE, THE DATASIZE OF SPD GENERATION AND VC TRAINING (IT SHOULD BE NOTED THAT THE VC TRAINING DATASIZE OF THE GROUP SYNTHETIC + NATURAL - NATURAL + SYNTHETIC IS TWICE THAT OF THE OTHER FOUR GROUPS RESPECTIVELY).

Speaker	Training data pair	TTS-450	TTS-400	TTS-200	TTS-80
Source - Target	Description	MCD / CER / WER	MCD CER WER	MCD / CER / WER	MCD / CER / WER
clb - slt	natural - natural	6.23 / 2.3 / 4.9	6.35 / 5.0 / 9.1	6.66 / 5.3 / 9.5	6.87 / 4.1 / 8.8
	natural - synthetic	**6.72 / 3.3 / 6.4**	**6.64 / 3.7 / 7.5**	**6.74 / 5.6 / 10.7**	**7.27 / 8.3 / 13.1**
	synthetic - natural	**6.74 / 4.3 / 8.0**	**6.68 / 3.4 / 6.7**	**6.68 / 3.1 / 6.5**	**6.96 / 5.7 / 11.5**
	synthetic - synthetic	6.77 / 5.3 / 8.2	6.85 / 4.8 / 8.5	6.97 / 8.5 / 13.1	8.33 / 19.6 / 25.6
	synthetic + natural - natural + synthetic	6.61 / 4.3 / 8.5	6.59 / 3.7 / 7.1	6.70 / 4.9 / 8.4	7.03 / 8.0 /12.6
bdl - rms	natural - natural	6.56 / 7.8 / 14.0	6.64 / 10.3 / 18.7	7.17 / 11.7 / 20.7	7.02 / 16.6 / 27.2
	natural - synthetic	**7.02 / 11.5 / 20.7**	**7.32 / 9.7 / 17.1**	**7.43 / 11.3 / 19.5**	**7.72 / 12.7 / 20.4**
	synthetic - natural	**6.63 / 10.2 / 18.5**	**6.81 / 9.4 / 16.4**	**6.94 / 10.7 / 18.6**	**7.19 / 11.4 / 18.8**
	synthetic - synthetic	7.07 / 10.2 / 18.8	7.36 / 8.1 / 15.3	7.51 / 11.6 / 20.3	7.82 / 14.4 / 24.8
	synthetic + natural - natural + synthetic	6.91 / 8.8 / 16.6	7.29 / 8.9 / 15.0	7.37 / 10.6 / 18.0	7.70 / 12.1 / 21.0
		Quality of synthetic data			
	Synthetic pair	MCD / CER / WER	MCD / CER / WER	MCD / CER / WER	MCD / CER / WER
	clb (synthetic)	6.40 / 3.3 / 6.0	6.44 / 3.5 / 6.4	6.67 / 3.1 / 5.7	7.31 / 4.3 / 6.3
	slt (synthetic)	6.32 / 4.1 / 6.8	6.32 / 3.7 / 6.5	6.40 / 4.5 / 7.7	7.01 / 8.6 / 13.1
	mean of MCD	**6.36**	**6.38**	**6.54**	**7.16**
	bdl (synthetic)	6.88 / 5.1 / 8.5	6.78 / 5.0 / 8.4	6.96 / 4.1 / 6.4	7.68 / 4.8 / 7.4
	rms (synthetic)	6.61 / 4.0 / 8.0	6.79 / 4.1 / 7.3	7.05 / 4.4 / 8.1	7.49 / 6.9 / 11.0
	mean of MCD	**6.75**	**6.79**	**7.00**	**7.59**

To generate high-quality synthetic data, we used the Parallel WaveGAN (PWG) neural vocoder [20], [21] to implement on the TTS and VTN model. PWG is a non-autoregressive vocoder, which allows parallel generation, and it is more efficient for the real-time waveform generation. Here we followed the open-source implementation[3]. In terms of the different target speakers we needed, the corresponding speaker-dependent PWG vocoders were trained by using the full dataset from CMU ARCTIC database.

We performed objective evaluations using several metrics, such as mel cepstrum distortion (MCD) to capture spectral envelope distortion between generated speech and natural speech, and word error rate (WER) and the character error rate (CER) to evaluate the intelligibility [8]. The ASR engine was Transformer-based model [22] which was trained by LibriSpeech database [23].

We also conducted subjective tests to evaluate VC perceptual performance from two perspectives, naturalness and speaker similarity of generated speech. In the naturalness test, an opinion test was conducted to evaluate naturalness with a mean opinion score (MOS). Listeners were asked to rate (in one-to-five scale) the naturalness of each given speech. In the speaker similarity test, a preference test was conducted. A pair of the converted speech and the ground truth speech were presented to the listeners at the same time. And they were asked to judge whether the two utterances produced by the same speaker on a four-point scale.

B. The investigation of feasibility and property on SPD

This part mainly focuses on the experimental investigation of Q1 from Section I. We further divide the Q1 into two sub-questions:

Q1-1: How does quality of data affect VC performance?

[3]https://github.com/kan-bayashi/ParallelWaveGAN

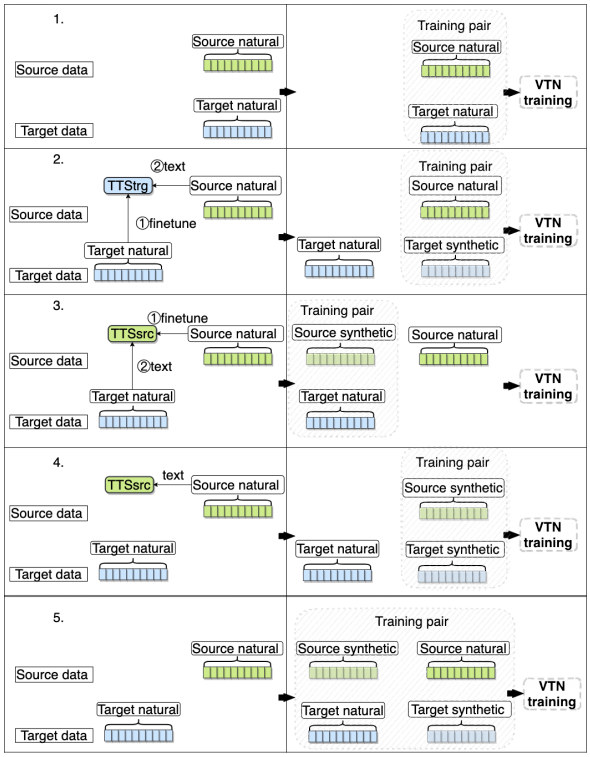

Fig. 2. Implementation method of five groups of experiments. It should be noted that the proportion of synthetic and natural corpus is the same in both source and target speakers in the group 5.

Q1-2: Which kind of the training pair is better?

For Q1-1, given that there is certain amount of non-parallel original training dataset, the TTS models need to be trained in advance using the non-parallel dataset to generate SPD.

TABLE II
EXPERIMENTAL RESULTS UNDER DIFFERENT SEMIPARALLEL SETTING.

Original datasize: 400					
Speaker Source	Speaker Target	Parallel ratio (%)	Training size	MCD / CER / WER	Eliminated training size	MCD / CER / WER
clb	slt	0	800-800	6.59 / 3.7 / 7.1	-	-
		25	700-700	6.61 / 4.5 / 9.2	400-400	6.85 / 2.9 / 5.0
		50	600-600	6.67 / 8.0 / 13.8	400-400	6.73 / 3.4 / 5.8
		75	500-500	6.64 / 2.4 / 4.9	400-400	6.74 / 3.3 / 6.4
		100	400-400	6.35 / 5.0 / 9.1	-	-
Original datasize: 40						
Speaker Source	Speaker Target	Parallel ratio (%)	Training size	MCD / CER / WER	Eliminated training size	MCD / CER / WER
clb	slt	0	80-80	7.69 / 12.2 / 20.6	-	-
		25	70-70	7.41 / 8.5 / 15.8	40-40	6.94 / 6.7 / 12.2
		50	60-60	7.07 / 5.7 / 10.1	40-40	6.81 / 6.0 / 12.6
		75	50-50	6.86 / 4.0 / 8.4	40-40	6.80 / 5.5 / 10.5
		100	40-40	6.80 / 4.1 / 8.9	-	-

Here, we use the term datasize to represent the amount of utterances. Therefore, the original training datasize will affect the TTS model performance and determine the total VC training volume, which can be taken as the two indicators of Q1-1. For Q1-2, it can be concluded from Fig. 1 that adding SPD will generate different types of training pairs. Hence, different training pairs may affect the VC results, which needs to be further investigated. Here, under the condition that the development set (100) and the evaluation set (100) remain unchanged, we choose the other utterances (932) of the CMU ARCTIC database to divide different quantities of non-parallel training sets for comparison experiments. The experimental process is shown in Fig. 2. Each experiment is implemented with 5 groups:

1. <source natural, target natural>
2. <source natural, target synthetic>
3. <source synthetic, target natural>
4. <source synthetic, target synthetic>
5. <source synthetic and source natural, target natural and target synthetic>

TABLE I lists the results from different sizes of training data. Female-speaker pair (clb, slt) and male-speaker pair (bdl, rms) are conducted to do VC training respectively. Here, mean of MCD represents the quality of synthetic data which also reflects the performance of corresponding TTS models. The overall quality of synthetic data produced by female speakers is better than that produced by male speakers in each datasize. This is because the original TTS model which is pretrianed by the data of female speaker (judy) in M-AILABS database, inherit the common characteristics of female speakers and have the better adaptation in synthesizing female speech.

The experimental results show that the VC results from using pure natural parallel data are always the best among the five groups. On the contrary, the VC results are the worst by only using SPD for training. From the overall trend, it is obvious that larger datasize leads to better TTS performance.

In general, under each training datasize, synthetic−natural tends to show the best VC results among the three training pairs using SPD. However, the difference of the VC results between natural−synthetic and synthetic−natural tends to be marginal as the datasize increases. As the datasize decreases, the gap between the results of synthetic-natural and natural-synthetic gradually increases. These results suggest that the source side is more robust against the quality of the synthetic data than the target side. Therefore, caution should be taken on the quality of the synthetic data to reduce the the negative impact when the performance of TTS drops.

In addition, the VC results of mixed training pairs (synthetic+natural−natural+synthetic) become the second best when the datasize of clb−slt is greater than or equal to 400. The reason is that the quality of synthesized speech is excellent enough, mixed training pairs have lager high-quality datasize. On the other hand, it can be observed from bdl−rms that when the datasize is 450, the VC result of the mixed training pair is still slightly worse than that of synthetic−natural, which is different from the result obtained from clb−slt pair with the same datasize, indicating that the performance of the TTS models is not sufficiently high.

It can be concluded that in general case, the TTS performance is most critical in terms of the impact on the VC results. For Q1-1, it can be clarified that when the original datasize is small, the quality of SPD will be degraded due to the limited performance of the TTS model. Especially when the target speaker uses SPD or the entire training set only contains SPD, the VC result is generally unsatisfactory. Conversely, we can appropriately reduce the constraints on the use of SPD when the SPD quality is good enough, and use it together with natural data to ensure that the VC training datasize is large enough to achieve better VC results. For Q1-2, the training dataset of source synthetic−target natural generally performs better. However, when the TTS performance is excellent enough, the mixed training pairs will bring the optimal results.

C. The investigation for semiparallel setting

This part contains the investigation and experimental results from the semiparallel setting based on Q2 in section I. Parallel ratio (PR) is used to represent the proportion of natural parallel corpus, so as to reflect the semi-parallel setting. We gradually

TABLE III
EXPERIMENTAL RESULTS OF ADDING EXTERNAL DATA WITH DIFFERENT DATASIZES. TTS-400 AND TTS-200 REPRESENT HOMOLOGOUS DATASIZE OF TTS FINETUNING.

Speaker	**External Synthetic Speech Quality**	**TTS-400**			
clb - slt	WER	Natural - Natural	Natural - Synthetic	Synthetic - Natural	Synthetic - Synthetic
External Datasize	Source / Target	MCD / CER / WER	MCD / CER / WER	MCD / CER / WER	MCD / CER / WER
0	/	**6.35 / 5.0 / 9.1**	**6.64 / 3.7 / 7.5**	**6.68 / 3.4 / 6.7**	**6.85 / 4.8 / 8.5**
1k	0.0 / 0.0	6.19 / 1.8 / 4.6	6.48 / 3.0 / 6.6	7.18 / 8.4 / 12.1	8.13 / 27.3 /31.4
2k	0.0 / 0.0	6.15 / 1.6 / 4.0	6.52 / 3.0 / 7.1	6.69 / 3.9 / 6.6	8.29 / 32.7 / 36.8
5k	0.9 / 0.9	**6.22 / 1.8 /4.4**	**6.50 / 2.9 / 6.3**	**7.12 / 8.4 / 11.8**	**7.35 / 12.3 / 15.5**
Speaker	**External Synthetic Speech Quality**	**TTS-200**			
clb - slt	WER	Natural - Natural	Natural - Synthetic	Synthetic - Natural	Synthetic - Synthetic
External Datasize	Source / Target	MCD / CER / WER	MCD / CER / WER	MCD / CER / WER	MCD / CER / WER
0	/	**6.66 / 5.3 / 9.5**	**6.74 / 5.6 / 10.7**	**6.68 / 3.1 / 6.5**	**6.97 / 8.5 / 13.1**
1k	0.0 / 0.0	6.46 / 2.2 /4.9	6.66 / 4.0 / 7.4	6.69 / 3.9 / 7.9	7.72 / 17.9 / 22.3
2k	0.0 / 0.0	6.50 / 2.1 / 5.0	6.66 / 3.3 / 7.4	6.83 / 3.4 / 6.4	7.67 / 18.3 / 21.8
5k	0.6 / 0.8	**6.40 / 3.0 / 6.3**	**6.65 / 3.9 / 8.5**	**7.01 / 6.1 / 10.7**	**7.78 / 19.6 / 24.8**

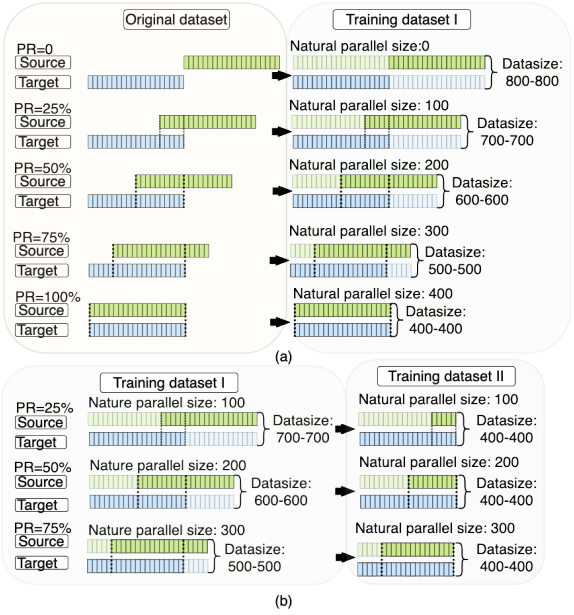

Fig. 3. Training procedure with different semiparallel setting (e.g., datasize=400). Training dataset I retains all SPD. Training dataset II removes natural−synthetic part.

increase the PR of training data pairs from totally non-parallel (PR=0%) to parallel (PR=100%) with the same datasize of the original training data. The respective TTS models of source speaker and target speaker are trained in case of constant datasize but different semiparallel setting for each group, which means the sum of datasize of VC parallel training involving synthetic data and natural data is varying. Different sizes of original training data are selected (datasize = 400, 40) to compare the VC results. Fig. 3 illustrates the procedure with the original datasize of 400. The experiments are divided into two parts: Using all feasible data for training, as shown in Fig. 3 (a); further removing the natural−synthetic part of semiparallel case for training, as shown in Fig. 3 (b).

The results of the experiments are listed in TABLE II. Without cutting training data, it can be observed that as PR is from 0% to 100%, the results become better when the original datasize is 40. In addition, except for the case of PR=1, the results with original datasize of 400 are relatively better at PR=0, but the overall trend does not change significantly. When the natural−synthetic part is removed, the semiparallel training results with the original datasize of 40 is improved, which is contrary to the original datasize of 400.

Based on the series of experiments, it can be concluded that the outcomes are related with the original training datasize:

(1) When training data level of TTS model is 400, the quality of synthetic data generated from TTS models is high. If PR = 0, the amount of VC training data is 800-800. Hence, the results can be improved by providing large datasize and high quality of synthetic data. With the increase of PR, the increase of the natural−natural part and the decrease of the synthetic datasize can compensate the negative effect caused by the decrease of the total training datasize. However, when the TTS performance is good enough, the datasize will become the critical factor. This also explains why further cancellation of the natural−synthetic part reduces VC performance.

(2) As the training datasize is small (e.g., 40), the resulting trained TTS models are unable to generate high quality data. Therefore, the conversion results should be poor when PR is low. Here, adjusting the training datasize while keeping the usage of synthetic data is unable to compensate the negative impact from the bad TTS performance. If PR and the ratio of using natural−natural are higher, the results will get better. Even so, the best results should not be better than any experimental result in which the datasize is 400, since the total amount of data is much less. Meanwhile, we find that in the case of semiparallel setting, eliminating the natural−synthetic part can improve the VC results, and with the increase of PR, the performance will be close to natural−natural (PR=100%) training. So natural−synthetic part has a great negative impact on VC when datasize is small, which is consistent with the conclusion based on Q1.

TABLE IV
RESULTS OF SUBJECTIVE EVALUATION USING TEST SET UNDER 450 AND 80 DATASIZE WITH 95% CONFIDENCE INTERVALS FOR Q1.

Speaker	**Training data**	**TTS-450**		**TTS-80**	
	Description	Naturalness	Similarity	Naturalness	Similarity
clb - slt bdl - rms	natural - natural	3.67 ± 0.14	71% ± 8%	3.42 ± 0.22	61% ± 8%
	natural - synthetic	3.43 ± 0.15	57% ± 8%	2.91 ± 0.21	46% ± 7%
	synthetic - natural	3.52 ± 0.14	62% ± 8%	3.26 ± 0.20	53% ± 9%
	synthetic - synthetic	3.33 ± 0.16	52% ± 7%	2.81 ± 0.21	43% ± 9%
	synthetic + natural - synthetic + natural	3.55 ± 0.14	63% ± 8%	3.12 ± 0.23	49% ± 8%

TABLE V
RESULTS OF SUBJECTIVE EVALUATION USING TEST SETS UNDER 400 AND 200 DATASIZE WITH 95% CONFIDENCE INTERVALS FOR Q3.

Speaker	**Training data**	**TTS-400**		**TTS-200**	
clb - slt	Description	Naturalness	Similarity	Naturalness	Similarity
non-exteral data	natural - natural	3.71 ± 0.11	73% ± 7%	3.69 ± 0.11	65% ± 6%
	natural - synthetic	3.45 ± 0.11	61% ± 7%	3.43 ± 0.15	53% ± 8%
	synthetic - natural	3.67 ± 0.12	69% ± 8%	3.54 ± 0.15	60% ± 7%
	synthetic - synthetic	3.32 ± 0.15	58% ± 7%	3.34 ± 0.14	51% ± 7%
adding-external data	natural - natural	4.10 ± 0.12	83% ± 6%	3.88 ± 0.13	69% ± 7%
	natural - synthetic	3.73 ± 0.14	70% ± 8%	3.55 ± 0.13	60% ± 7%
	synthetic - natural	3.57 ± 0.15	63% ±8%	3.38 ± 0.12	51% ± 8%
	synthetic - synthetic	3.10 ± 0.18	52% ± 9%	2.80 ± 0.17	40% ± 8%

TABLE VI
RESULTS OF SUBJECTIVE EVALUATION USING TEST SETS WITH 95% CONFIDENCE INTERVALS FOR Q2.

Speaker	**Training data**	**TTS-40**	
	Description	Naturalness	Similarity
clb - slt	PR = 0	1.95 ± 0.17	20% ± 7%
	PR = 50%	2.74 ± 0.19	47% ± 8%
	PR = 50% without natural-synthetic	3.68 ± 0.14	68% ± 7%
	PR = 100%	3.91 ± 0.12	78% ± 8%

D. The investigation on the external text data

This part is the experimental study to verify the influence of external text data corresponding to the Q3 in section I. We fix the original non-parallel dataset and train the TTS models. Then input all the external text data from M-AILABS database (15,369 utterances in total) to TTS models to generate the external SPD. WER is used as the basis for selecting the highest quality SPD. The external SPD with different sizes is input into four non-parallel training cases.

TABLE III presents the results of different original datasizes. By comparing the VC results of non-external data, we can find that the introduction of external data has positive impact on VC. Especially when the original training pair is natural−natural, introducing external data will significantly improve the VC results. On the other hand, the natural−synthetic pair also gets the better VC results after adding the external synthetic data. With more external data involved in the training, the result tends to be better. Nevertheless, the influence of external data on the synthetic−natural pair shows an opposite trend to former. As a result, it can be concluded that natural−synthetic outperforms the synthetic−natural after adding the external data.

Finally, the addition of external data is detrimental when original data is synthetic−synthetic. In other words, synthetic data incompletely inherits all the features of natural speech. On the contrary, natural data plays a corrective role. When natural part is lost, the VC model will fully learn the features of synthetic data, and external data enhances this learning process, leading to a worse effect.

Therefore, the composition of the original training dataset should be considered before introducing the external data. The original training datasets which are pure natural parallel or natural−synthetic can benefit from the external data. On the contrary, it is unnecessary to introduce external SPD under the circumstance that the original training pair contains source synthetic data.

E. Subjective evaluation

In this part, we take samples generated by experiments of the objective evaluation to construct the test sets for subjective evaluation tests.

For the subjective tests Q1, we feed the synthetic speech of male and female target speakers (slt, rms) into the test set. According to the naturalness and similarity results presented in TABLE IV, using synthetic−natural dataset is slightly better than that of using natural−synthetic under the datasize of 450. And this difference of the performance becomes significant when datasize is 80. Meanwhile, the performance of the method by using mixed training pair (synthetic+natural−natural+synthetic) is comparable with using synthetic−natural in the datasize of 450, but slightly worse in the datasize of 80.

For the subjective tests on Q2, the naturalness and similarity results are shown in TABLE VI. The evaluation of the performance is able to approach to PR=1 by increasing PR and removing synthetic−natural part of the dataset simultaneously under the semiparallel setting with small datasize.

For the subjective tests on Q3, the results of subjective test with datasize of 400 and 200 show a synchronous trend in TABLE V. It can be found that external SPD can improve the

performance of natural−synthetic and natural−natural.

The overall results are consistent with the findings in the objective evaluations. The answers to the three questions are summarized as follows:

A1: SPD is feasible for seq2seq non-parallel VC. SPD is produced by the TTS models trained with the original dataset of source and target speakers. Therefore, the VC results using SPD are determined by the performance of TTS models and VC training datasize. When the original data is sufficient, we can obtain the TTS models with excellent performance, resulting in a better VC result. In addition, the VC result is also affected by the object of using SPD. When the dataset is limited, providing SPD for the source speaker and retaining parallel natural data for the target speaker can get better VC result.

A2: When the dataset is semiparallel, we should try to ensure the PR is large enough. Under this premise, when the original datasize is large, the introduction of SPD into target speaker or source speaker can both achieve ideal VC results. Thus, the full use of all types of SPD to ensure amount of data, can maximize the benefits. On the contrary, when the original datasize is small, the well-performing TTS models are difficult to get. Introducing training pair with negative impact such as source natural−target synthetic should be avoided.

A3: The introduction of external text data can provide a large amount of useful parallel data for VC. External data can significantly improve the VC results when the original dataset is natural−natural and natural−synthetic. However, it should be noted that when there is no natural speech or only source natural speech in the original dataset, the introduction of external data will lead a negative impact on VC training.

IV. Conclusion

In this paper, a series of experiments are carried out to study the impact of using the SPD on non-parallel seq2seq VC and to address the three questions posed in Section I. The experimental results provide guidance for using synthetic speech. The results show that SPD is feasible in the absence of natural parallel data, and the VC results are related to both the TTS performance and the VC training datasize. When the original datasize is larger, the effect of using SPD is better. Generally, the VC results will benefit more by exclusively providing synthetic data to the source speaker than to the target speaker. However, this situation is reversed when the external data is added. Moreover, although the research systematically explores and verifies the different cases of SPD on seq2seq VC, we mainly focus on a limited number of speaker pairs. In addition, the maximum of the original datasize is only 450, which means that we are unable to determine the upper limit of the training effect that SPD can achieve for the time being. Therefore, using more speakers and a larger amount of data to investigate the beneficial trend that seq2seq non-parallel VC can obtain from SPD is the future research direction. In terms of methodology, we can introduce the the VC models which can directly processing non-parallel data to compare the performance with the way of using SPD on seq2seq VC in the future research, so as to further clarify the role of SPD.

V. Acknowledgement

This work was partly supported by JST CREST Grant Number JPMJCR19A3, and AMED under Grant Number JP21dk0310114, Japan.

References

[1] D. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," *1985 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, pp. 748-751, 1985.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131-142, 1998.

[3] 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, pp. 2222-2235, 2007.

[4] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Computer Science*, 2014.

[7] W. C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading ASR and TTS," *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 160–164, 2020.

[8] W. C. Huang, Y. C. Wu, T. Hayashi, and T. Toda, "Any-to-One Sequence-to-Sequence Voice Conversion Using Self-Supervised Discrete Speech Representations," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[9] J. X. Zhang, Z. H. Ling, and L. R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 540-552, 2019.

[10] Y. Zhang, H. Che, and X. Wang, "Non-parallel Sequence-to-Sequence Voice Conversion for Arbitrary Speakers," *2021 IEEE 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1-5, 2021.

[11] W. C. Huang, T. Hayashi, Y. C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," *Proc. Interspeech*, pp. 4676-4680, 2020.

[12] W. C. Huang, P. L. Tobing, Y. C. Wu, K. Kobayashi, and T. Toda, "The NU voice conversion system for the Voice Conversion Challenge 2020: On the effectiveness of sequence-to-sequence models and autoregressive neural vocoders," *ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020.

[13] S. H. Mohammadi, A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol.88, pp. 65-82, 2017.

[14] H. Duxans, D. Erro, J. Pérez, F. Diego, A. Bonafonte, and A. Moreno, "Voice conversion of non-aligned data using unit selection," *TC-STAR WSST*, 2006.

[15] J. Kominek, A. W. Black, "The CMU Arctic speech databases," *Fifth ISCA workshop on speech synthesis*, 2004.

[16] Munich Artificial Intelligence Laboratories GmbH, "The MAILABS speech dataset," 2019, accessed 30 November 2019.

[17] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[18] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," *Proc. INTERSPEECH*, pp. 2207-2211, 2018.

[19] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network." *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6706-6713, 2019.

[20] R. Yamamoto, E. Song, and J. M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199-6203, 2020.

[21] P. L. Tobing, Y. C. Wu, and T. Toda, "Baseline system of Voice Conversion Challenge 2020 with cyclic variational autoencoder and Parallel WaveGAN," *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 155-159, 2020.

[22] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884-5888, 2018.

[23] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206-5210, 2015.