ドメイン適応と話者一致損失を用いた話者適応による クロスリンガル音声合成*

☆辛 徳泰，齋藤 佑樹，高道 慎之介，郡山 知樹，猿渡 洋（東大院・情報理工）

1 Introduction

Cross-lingual text-to-speech (TTS) synthesis refers to a task that requires the TTS model to synthesize a source language's speech for a target language's speaker. Previous work about cross-lingual TTS usually trains a multilingual multi-speaker TTS model on multiple monolingual datasets by conditioning the TTS model on speaker and language representations called speaker and language embeddings, respectively [1, 2]. One of the drawbacks of such a method is that it requires a large amount of speech data with transcriptions for speakers of the target language. However, in practice not necessarily all speakers of the target language have such an amount of data. In such a case speaker adaptation can be used, which refers to a technology that can adapt the TTS model to a new speaker whose data is not included in the training set and enable the TTS model to synthesize the target speaker's speech at a relatively lower cost. Existing monolingual speaker adaptation methods often use a small amount of the target speaker's speech data to fine-tune a pretrained multi-speaker TTS model [3]. However, since the TTS model of the source language never sees the target language's data during the training process, this method is not applicable for cross-lingual speakers, i.e., speakers of different languages.

In this paper, we propose a cross-lingual speaker adaptation method using domain adaptation and speaker consistency loss for TTS synthesis. The basic idea of our work is inspired by Jia et al. [4], a monolingual speaker adaptation method. It only needs multilingual data without transcriptions to train a speaker verification model and the source language's data to train a TTS model. Also, the proposed method can fine-tune the source language's TTS model on the target language's data by using the speaker embeddings of the target speakers, which only needs up to 3 minutes of speech data of each speaker. Inspired by our previous work [1, 2], the proposed method first trains a language-independent speaker verification model on multilingual data by using domain adaptation. Then the speaker embeddings for source language's speakers are used to train a monolingual multi-speaker TTS model. To adapt the TTS model to cross-lingual target speakers, the proposed method fine-tunes the TTS model under a speaker consistency loss that maximizes the cosine similarity between speaker embeddings of the same speaker generated from the target language's natural speech and the source language's synthesized speech. Experimental results demonstrated that (1) the proposed method can synthesize speech of the source language of a cross-lingual speaker with higher speech naturalness than the conventional method; (2) using speaker consistency loss with a language-dependent speaker verification model will instead causing performance degradation. The audio samples of this work are publicated on our project page: `https://aria-k-alethia.github.io/2021clsa/`.

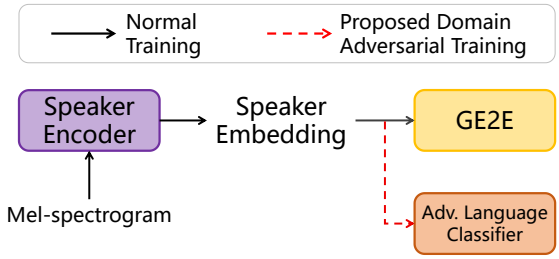

Fig. 1 Architecture of the speaker encoder. The adversarial language classifier is not used in the conventional method.

2 Baseline Method

In this section we introduce a baseline method for cross-lingual speaker adaptation based on previous work [4] used in the experiments. The baseline method contains two components: (1) a speaker encoder based on multilingual speaker verification, (2) a monolingual multi-speaker TTS model based on Tacotron2.

2.1 Speaker Encoder based on Speaker Verification

Recent work has shown that speaker verification/recognition is an appropriate pretext task to learn speaker embeddings for multi-speaker TTS [1, 5]. This is because learning a representative and distinct speaker embedding for each speaker is a common requirement for speaker verification and multi-speaker TTS. Thus in this stage we train the speaker encoder based on speaker verification using multilingual data.

The general architecture of the speaker encoder is

*Cross-lingual Speaker Adaptation using Domain Adaptation and Speaker Consistency Loss for Text-To-Speech Synthesis, by Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, Hiroshi Saruwatari (The University of Tokyo).

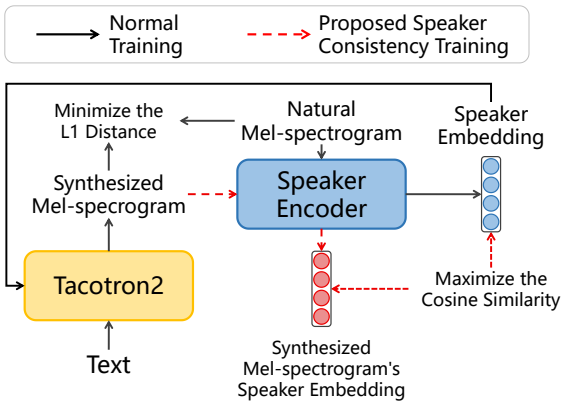

Fig. 2 Diagram of the proposed TTS model and the speaker consistency loss.

illustrated in Figure 1. Basically it is similar to the architecture mentioned in Jia et al. [4]. The only difference here is that we use a more powerful model called ResCNN [6] as the speaker encoder's implementation instead of simple long short-term memory to encode the mel-spectrograms to cope with the complexity of multilingual data. We use generalized end-to-end (GE2E) loss $\mathcal{L}_{ge2e}$ [7] as the speaker verification loss.

2.2 Multi-speaker Tacotron

After we get the trained speaker encoder, we train a monolingual multi-speaker TTS model by conditioning a Tacotron2-based TTS model [8] on the speaker embedding generated by the speaker encoder (Figure 2). Specifically, following previous work [4] we concatenate the speaker embedding to the output states of the text encoder and use this as the input of the attention module to let the decoder attend over them. In this stage we only use the source language's data and keep the parameters of the speaker encoder fixed.

Formally, denote the mel-spectrogram of the jth utterance of the ith speaker as x_{ij}, the embedding generated by speaker encoder $f_s(\cdot)$ is defined as: $e_{ij} = f_s(x_{ij})$. Denote the Tacotron2 model and the text sequence of the mel-spectrogram x_{ij} as $f_t(\cdot)$ and y_{ij}, respectively, the loss function of the TTS model is defined as the L1 distance between the natural mel-spectrogram x_{ij} and the synthesized one $\tilde{x}_{ij} = f_t(y_{ij}, e_{ij})$: $\mathcal{L}_{tts} = \sum_{i,j} |\tilde{x}_{ij} - x_{ij}|$. After training, to adapt to a new speaker, we first average speaker embeddings generated from the speaker's k utterances and use the averaged speaker embedding to synthesize the source language's speech.

3 Proposed Method

In this section we describe the proposed method by extending the baseline method described in Section 2. The proposed method first trains a speaker encoder based on speaker verification using GE2E loss and domain adaptation to construct a language-independent speaker space. To fine-tune the TTS model on cross-lingual speech, the proposed method then uses a speaker consistency loss to maximize the cosine similarity between the speaker embeddings of the natural and the same speaker's synthesized mel-spectrograms, which can benefit from the language-independent speaker space learned by the domain adaptation.

3.1 Language-independent Speaker Verification based on Domain Adaptation

The baseline speaker verification method introduced previously (Section 2.1) is suitable for monolingual settings. However, when training the model on multiple datasets of different languages, the language-dependent speakers cause a domain shift between the speakers of different languages [9]. This impedes the knowledge transferring from the speaker encoder to the TTS model for cross-lingual speakers. Therefore to make it possible to fine-tune the TTS model on cross-lingual data, following our previous work [1, 2] we use a domain adaptation algorithm to eliminate the domain shift.

Specifically, we use domain adversarial neural network (DANN) [10]. As illustrated in Figure 1, the proposed method extends the baseline method by adding an adversarial language classifier. During training the classifier tries to learn the language label from the speaker embedding. Meanwhile, the adversarial training forces the model to exclude information relating to language from the speaker embedding, which finally makes the speaker encoder become language-independent. This is accomplished by a gradient reversal layer (GRL) to reverse the gradient back-propagated from the classifier to the speaker embedding.

Formally, denote the GRL operator, the adversarial language classifier and the language label of x_{ij} as $\triangle$, $f_l(\cdot)$ and l_{ij}, respectively, the loss of DANN is defined as:

$$\begin{aligned}\mathcal{L}_{da} = \sum_{i,j} & -l_{ij} \log \sigma(f_l(\triangle e_{ij})) \\ & - (1 - l_i) \log (1 - \sigma(f_l(\triangle e_{ij}))),\end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. The final loss function of the proposed speaker encoder is defined as: $\mathcal{L}_s = \mathcal{L}_{ge2e} + \mathcal{L}_{da}$. This loss makes the speaker space generated by the speaker encoder not only reflect the speaker characteristics but also not depend on language.

3.2 Speaker Adaptation based on Consistency Loss

As mentioned previously cross-lingual speech data can not be used to fine-tune a monolingual TTS model of source language directly. However, the knowledge learned by the pretrained speaker encoder can be utilized. Thus we propose a speaker

consistency loss (SCL) to maximize the cosine similarity between speaker embeddings extracted from the natural and the same speaker's synthesized mel-spectrogram. The core idea is inspired by Nachmani et al. [11], which has shown a possible way to apply such loss to the cross-lingual TTS task. While their work only used the gradient of the loss to update the speaker encoder's parameters, which has little influence on the speech synthesis process, in this work we used the gradient to update the parameters of the decoder in the Tacotron2 model.

As illustrated in Figure 2, to compute the SCL the proposed method first feeds the text and the speaker embedding extracted from a target speaker's natural mel-spectrogram to the multi-speaker Tacotron2 model to synthesize a mel-spectrogram of the source language. The synthesized mel-spectrogram is then fed to the speaker encoder to get the speaker embedding of the synthesized mel-spectrogram of the same speaker. Finally the cosine similarity between the two embeddings of the same speaker is maximized. In the fine-tuning stage we froze the parameters of the speaker encoder and the Tacotron2 model except for the mel-spectrogram decoder. Formally, the SCL is defined as:$\mathcal{L}_{scl} = -\sum_{i,j} \cos\left(f_s(\tilde{x}_{ij}), e_{ij}\right))$. In addition, during experiments we found only using cross-lingual data to fine-tune would decrease the speech quality. We then computed SCL for both intra-lingual and cross-lingual speakers to fine-tune the TTS model and saw improvements in this case. To stabilize the fine-tuning, we still include the L1 loss in this process. Thus the loss function for fine-tuning is defined as: $\mathcal{L}_{ft} = \mathcal{L}_{tts} + \alpha\mathcal{L}_{scl}$, where α is a hyperparameter.

4 Experiments

4.1 Experimental Setup

In our experiments, English is regarded as the source language, and Japanese is regarded as the target language. We used English multi-speaker dataset VCTK [12] and Japanese multi-speaker dataset JVS [13] to train the speaker verification model. The total number of speakers was 207 in which 107 were English speakers and 100 were Japanese speakers. We randomly picked up 8 English speakers and 8 Japanese speakers (both contained 4 female and 4 male speakers) as unseen speakers and excluded their data from the training and fine-tuning process. All audios were downsampled to 16 kHz and converted to 80-dimensional mel-spectrograms. The frame number of the mel-spectrogram was segmented to $[120, 150]$ for training and inference of the speaker verification model. After training the speaker encoder we trained the TTS model by the L1 loss $\mathcal{L}_{tts}$ using the VCTK dataset. Then we fine-tuned the TTS model using the fine-tuning loss $\mathcal{L}_{ft}$ with both English and Japanese data. Finally we used WaveRNN [14] to convert the synthesized mel-spectrogram to the time-domain waveform.

In all experiments, the dimension of the speaker embedding was set to 64. The number of utterances k for computing speaker embedding and the weight hyperparameter α of SCL were empirically set to 5 and 0.1, respectively. We multiplied the gradient back-propagated from language classifier to the embedding by a factor $\lambda_p = \frac{2}{1+\exp(-10 \cdot p)} - 1$, where p represents the training progress ranging from 0 to 1.

We trained four models for comparison. The first two models were trained using the baseline and the proposed methods (Section 2, 3), which are denoted by Base. and DA+SCL, respectively. For ablation experiments we additionally trained two models. The Base.+SCL model was trained by using SCL with baseline speaker encoder but without domain adaptation. The DA model was trained by the proposed method without SCL. Note that we didn't fine-tune the TTS model for the two models without using SCL (Base. and DA).

4.2 Subjective Evaluation

We evaluated the synthesized speech from two aspects: speech naturalness and speaker similarity. In addition to the 16 unseen speakers, we randomly picked 16 speakers from the training set as seen speakers. Note that, since models without using SCL have no fine-tuning step, the 8 Japanese seen speakers are seen by the speaker encoder but completely unseen by the TTS model of Base. and DA. For each speaker we synthesized 20 utterances. All of these utterances were not included in the training process. We used Amazon Mechanical Turk[1] to conduct the subjective evaluation.

4.2.1 Speech Naturalness

We conducted 5-scale mean opinion score (MOS) tests to evaluate the speech naturalness of the synthesized speech. We categorized all utterances into four groups by their speaker (seen/unseen) and task (intra-lingual/cross-lingual); the MOS was calculated separately for each group. Totally 720 English listeners joined in the evaluation; each test had 90 listeners with 25 answers per listener. The result is shown in Table 2, in which ground truth (GT) represents the natural speech. It can be observed that the proposed DA+SCL model obtains the best performance in all tasks. For the intra-lingual task, Base.+SCL also obtains significant improvements compared to Base. and DA, demonstrating that the SCL can improve the intra-lingual speaker adaptation. In contrast, for the cross-lingual tasks the performance of Base.+SCL degrades significantly while the DA+SCL model still maintained similar performance compared to their performance on the intra-

[1] https://www.mturk.com/

Table 1 Results of XAB tests on speaker similarity. **Bold** scores indicate preferred method has p value less than 0.05

Task	Speaker	Base. vs. DA	Base.+SCL vs. DA+SCL	DA vs. DA+SCL	Base.+SCL vs. DA
Intra-lingual	Seen	0.464 - **0.536**	0.473 - **0.527**	0.500 - 0.500	0.468 - **0.532**
	Unseen	0.450 - **0.550**	0.448 - **0.552**	0.479 - **0.521**	0.504 - 0.496
Cross-lingual	Seen	0.440 - **0.560**	0.408 - **0.592**	0.492 - 0.508	0.408 - **0.592**
	Unseen	**0.524** - 0.476	**0.531** - 0.469	0.488 - 0.512	0.496 - 0.504

Table 2 Results of MOS evaluation on naturalness. **Bold** indicates better method comparing to Base. without overlapping 95% confidence interval

Task	Spkr.	Base.	Base.+ SCL	DA	DA+ SCL	GT
Intra	Seen	3.51	**3.65**	3.58	**3.67**	4.04
	Unseen	3.62	**3.73**	3.61	**3.77**	4.15
Cross	Seen	3.39	2.84	**3.60**	**3.61**	4.04
	Unseen	3.54	3.07	3.48	**3.55**	4.06

lingual tasks. This demonstrates that the domain shift makes it difficult to apply SCL directly for adapting the TTS model to cross-lingual speakers. Finally the DA model also has comparable performance to the DA+SCL model, which shows the effectiveness of removing the domain shift for the cross-lingual tasks.

4.2.2 Speaker Similarity

We then conducted preference XAB tests to evaluate the speaker similarity of the synthesized speech. Here X is the natural speech, which is English for VCTK speakers and Japanese for JVS speakers. We chose four pairs among all six combinations of the four models. Totally 800 listeners joined in the evaluation; each test had 25 listeners with 10 answers per listener. The result is shown in Table 1. The DA and DA+SCL models both obtain relatively better performance than the Base. and Base.+SCL models, respectively, which shows the effectiveness of the proposed method. Surprisingly, the performance of DA and DA+SCL are almost the same, which implies the SCL has relatively little influence on speaker similarity.

5 Conclusions

This paper described a method for cross-lingual speaker adaptation for TTS synthesis based on a language-independent speaker encoder using domain adaption and a speaker consistency loss which makes it possible to fine-tune the TTS model of source language on the target language's data. Experimental results demonstrated that the proposed model could significantly improve speech naturalness compared to the baseline method.

謝辞：本研究は，JSPS 科研費 19H01116, 18K18100, 17H06101, JSPS-CAS 二国間交流事業 JPJSBP120197203 による支援を受けたものです。

参考文献

[1] D. Xin et al., "Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space," in *Proc. Interspeech*, Shanghai, China, Oct. 2020.

[2] ——, "Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual TTS," in *Proc. ICASSP*, Toronto, Canada, Jun. 2021.

[3] H. B. Moss et al., "Boffin TTS: Few-shot speaker adaptation by bayesian optimization," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 7639–7643.

[4] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[5] J. Cho et al., "Learning speaker embedding from text-to-speech," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 3256–3260.

[6] C. Li et al., "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, May 2017.

[7] L. Wan et al., "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*, Calgary, Alberta, Canada, Apr. 2018, pp. 4879–4883.

[8] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Alberta, Canada, Apr. 2018, pp. 4779–4783.

[9] W. Xia et al., "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 5816–5820.

[10] Y. Ganin et al., "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] E. Nachmani, L. Wolf, "Unsupervised polyglot text-to-speech," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 7055–7059.

[12] C. Veaux et al., "Superseded-CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[13] S. Takamichi et al., "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[14] N. Kalchbrenner et al., "Efficient neural audio synthesis," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 2410–2419.