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

Cross-lingual Text-To-Speech Synthesis via Domain Adaptation and Perceptual Similarity Regression in Speaker Space

Detai Xin, Yuki Saito, Shinnosuke Takamichi, Tomoki Koriyama, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo, Japan

{detai_xin, yuuki_saito, shinnosuke_takamichi, tomoki_koriyama, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

Abstract

We present a method for improving the performance of cross-lingual text-to-speech synthesis. Previous works are able to model speaker individuality in speaker space via speaker encoder but suffer from performance decreasing when synthesizing cross-lingual speech. This is because the speaker space formed by all speaker embeddings is completely language-dependent. In order to construct a language-independent speaker space, we regard cross-lingual speech synthesis as a domain adaptation problem and propose a training method to let the speaker encoder adapt speaker embedding of different languages into the same space. Furthermore, to improve speaker individuality and construct a human-interpretable speaker space, we propose a regression method to construct perceptually correlated speaker space. Experimental result demonstrates that our method could not only improve the performance of both cross-lingual and intra-lingual speech but also find perceptually similar speakers beyond languages.

Index Terms: text-to-speech, cross-lingual, domain adaptation, speaker embedding

1. Introduction

Recent advances in speaker individuality modeling based on deep neural network (DNN) [1] have made multi-speaker end-to-end text-to-speech (TTS) synthesis [2, 3, 4, 5] become possible by conditioning a TTS model on a distributed representation which is usually called speaker embedding. Multilingual speech synthesis by a single TTS model is also possible by utilizing language feature [6, 7] and unified text representation like phoneme [8] or byte [9]. By combining the techniques mentioned above, it is possible to synthesize multilingual speech of one speaker even if there is only monolingual data of the speaker. This process is called cross-lingual TTS synthesis.

However, given a source language speaker, how to maintain the naturalness and speaker similarity of the speech when synthesizing target language speech is still an unsolved problem. Although it is assumed that the speaker embedding extracted from acoustic features encodes general voice information of the speaker, in practice embeddings of different languages usually gather in different clusters, which implies the speaker embedding is language-dependent. In this paper, we refer this problem as the language-dependent problem of speaker embedding in cross-lingual TTS synthesis.

One way to alleviate this problem is introducing common text representation such as phoneme or byte. Chen et al. [10] and Zhang et al. [8] proposed cross-lingual TTS model based on Tacotron2 [11] by using this method. Although they found phoneme was suitable for this task, the performance decreased when doing cross-lingual TTS synthesis, and the language-dependent problem remained.

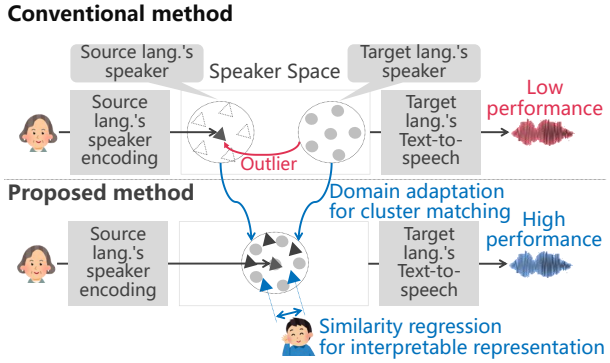


Figure 1: *Overview of the core idea of our work. With domain adaptation, speakers in source languages, which are outliers of those in target languages, can transfer their voices to target languages. With speaker similarity regression, speakers' voices are placed to be strongly correlated with humans' perception. Lang. indicates language.*

Another way to resolve the discrepancy is adding restrictions on the speaker space directly. Nachmani et al. [12] proposed a novel speaker preserving loss term to solve the problem. They assumed the cross-lingual speech synthesized by a pre-trained multilingual TTS model was ground truth data and minimized the L1 distance between speaker embedding of ground truth intra-lingual speech and synthesized cross-lingual speech. However, this method had a bad influence on the quality of synthesized speech, since the synthesized speech was treated as ground truth data in the training process. Most recently, Maiti et al. [13] proposed a semi-supervised algorithm that could transfer any source speaker embedding to target speaker space by computing the difference between the speaker embedding of source and target language of a bilingual speaker. This demonstrated better performance than the baseline model that didn't transfer speaker embedding. However, it is not practical to apply this method to models with multiple languages, since seeking a speaker who can speak all languages is almost impossible.

In this paper, we propose a method for resolving language dependency in speaker space by regarding cross-lingual TTS synthesis as a domain adaptation problem. As illustrated in Figure 1, when synthesizing target language speech of a source speaker, the voice information of the source speaker can be encoded to the language-independent space to assist the synthesis and consequently improves naturalness and similarity. Our idea is inspired by the domain adversarial neural network (DANN) [14], which has been proved to be an effective algorithm for domain adaptation. An adversarial loss term is added in our speaker encoder to force it to ignore the difference between languages. Furthermore, to improve speaker individuality of the synthesized speech and make the speaker space constructed

http://dx.doi.org/10.21437/Interspeech.2020-2070

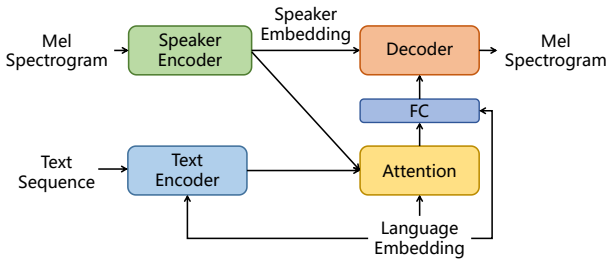


Figure 2: *General architecture of cross-lingual TTS synthesis. Here FC refers to fully connected layer.*

by our method correlates to human perception, we use inter-speaker perceptual similarity score [15] as an additional loss term. Experimental results demonstrate that (1) our method improves speech naturalness in not only cross-lingual but also intra-lingual speech synthesis, (2) our model can find speakers with similar voice, even they speak different languages. Audio samples of our work are publicated on our project page: `https://aria-k-alethia.github.io/clttsda/`.

2. Conventional Method

In this section, we introduce a conventional method of cross-lingual TTS synthesis [8, 10, 12, 13]. Figure 2 illustrates the general architecture of the method. Generally, it consists of two parts: speech synthesis part and speaker encoder. The speech synthesis part is extended from Tacotron2 [11] but further conditions on speaker embedding and language embedding. The speaker encoder is used to extract a distributed representation of the speaker called speaker embedding from mel-spectrogram. Usually, DNN is used as the architecture of the speaker encoder. Besides, each language has a randomly initialized trainable embedding as a hidden variable to model speaker-independent language feature. The language embedding has been proved to be effective in improving the naturalness of speech synthesized by multilingual TTS model [8]. The cross-lingual TTS model is first trained with multilingual and multi-speaker speech data. When synthesizing cross-lingual speech of a source language speaker, target language text and embedding are fed to the model, and the speaker embedding is obtained from source language mel-spectrogram. We basically follow the original architectures but slightly modified it to improve the performances. The modification is used for both this conventional and our proposed method. See Section 4 for the detail.

One may assume the speaker embedding should only encode language-independent voice feature, but the speaker embedding generated by a conventional method usually forms different clusters according to different languages. This implies the speaker embedding encodes not only general voice feature but also language-dependent feature, which makes any source speaker embedding an outlier when synthesizing target language speech.

3. Cross-lingual TTS via Domain Adaptation and Perceptual Similarity

To solve the language dependency problem, we extend the speaker encoder of conventional method illustrated in Figure 2. We propose (1) domain adaptation objective (Section 3.1) and (2) inter-speaker similarity regression objective in speaker space (Section 3.2). The general architecture of our speaker encoder is illustrated in Figure 3.

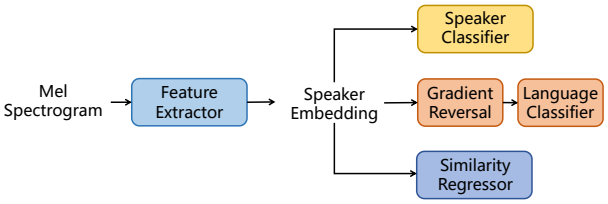


Figure 3: *Speaker encoder architecture of proposed cross-lingual TTS model.*

3.1. The Domain Adaptation Objective

In order to obtain language-independent speaker embedding, we regard the speaker classifier and language classifier as label predictor and domain classifier in DANN [14], respectively. The domain classifier in our work tries to minimize the language classification loss. On the other hand, the gradient reversal layer (GRL) reverses the gradient before it is propagated to the feature extractor, which actually makes the feature extractor have an opposite objective of domain classifier, i.e., ignoring the language difference of each speaker. Meanwhile, the speaker classifier makes the feature extractor learn text-independent voice feature from the input, which forms a minimax game with the domain classifier. After converging at a saddle point, the extracted speaker embedding will not include language information but still retain speaker identity information. As a result, all speaker embeddings together form a language-independent speaker space. The architecture of the feature extractor is based on resCNN [16], which is an effective deep residual convolution neural network for speaker identification task. We find resCNN is useful for extracting text-independent feature when we use adversarial loss, while other simple architectures like long short-term memory (LSTM) fails to converge.

Formally, we denote feature extractor, speaker classifier and language classifier as G_f, G_y and G_d respectively. For speaker i, given the mel-spectrogram x_i and corresponding speaker embedding $e_i = G_f(x_i)$, the one-hot speaker label vector y_i, the binary language label l_i, the domain adaptation loss $\mathcal{L}_i^{\text{DA}}$ for this speaker is computed by summing the classification loss of the two classifiers:

$$\mathcal{L}_i^{\text{DA}} = -\sum_{j=1}^{N_s} y_i^{(j)} \log \text{softmax}_j(G_y(e_i)) - l_i \log \sigma(G_d(\triangle e_i)) - (1 - l_i) \log\left(1 - \sigma(G_d(\triangle e_i))\right), \quad (1)$$

where N_s is the number of speaker, σ and $\triangle$ are the sigmoid function and gradient reversal operation, respectively. We use binary classification loss since the dataset used in experiments only contains two languages.

3.2. The Inter-speaker Similarity Regression Objective

Our domain adaptation method makes it possible to compare the similarity of speakers even they speak different languages. However, it has no guarantee that the speaker space generated by the speaker encoder correlates with subjective inter-speaker similarity, i.e., speakers with similar voice are not necessarily close to each other. Thus, to further improve the speaker individuality of the synthesized speech and construct a human-interpretable speaker space (i.e., highly correlated with human perception), we use inter-speaker perceptual similarity [15] to train the speaker encoder. This score is obtained by crowdsourcing involving a large number of human evaluators. Each evaluator is asked to give a preference score of voice similarity of two speakers. Thus the scores could represent the percep-

Table 1: *Results of MOS evaluation on naturalness. The language column represents the language of synthesized speeches.* ***Bold*** *indicates better method without overlapping* 95% *confidence interval*

Task	Language	Speaker	Conv.	Conv.+sim	DA	DA+sim	Ground truth
Intra-lingual	English	Seen	3.18	3.24	**3.25**	**3.27**	4.07
		Unseen	3.19	**3.26**	**3.24**	**3.24**	4.00
	Japanese	Seen	2.71	**2.80**	**2.80**	2.73	4.44
		Unseen	2.88	**2.95**	2.89	2.64	4.35
Cross-lingual	English	Seen	3.19	3.23	**3.36**	**3.34**	3.82
		Unseen	3.25	3.23	**3.31**	3.14	3.95
	Japanese	Seen	2.68	**2.84**	2.79	**2.80**	4.33
		Unseen	2.64	2.78	**2.98**	**2.94**	4.42

tually correlating level of two speakers. However, since it is originally used to construct human-interpretable speaker space for monolingual TTS system, the scores are only provided for intra-lingual speakers. The straightforward way to apply it to cross-lingual TTS system is to annotate scores for cross-lingual speaker pairs. But this is not realistic due to the scalability of the number of languages. In this paper, we solve this problem by combining speaker similarity regression and domain adaptation. The speaker similarity regression works to construct human-interpretable but language-specific speaker spaces. Meanwhile, the domain adaptation is used to match these spaces. Therefore, the intra-lingual similarity knowledge can be easily transferred to cross-lingual speaker pairs.

We normalize ground truth score between $[-1, 1]$, and augment Equation (1) with a regression loss term. Given the inter-speaker similarity score $s_{i,j}$ for speaker i and j, the final loss value of speaker i is:

$$\mathcal{L}_i^{\text{sim}} = \mathcal{L}_i^{\text{DA}} + \sum_{j \in \{k | l_k = l_i\}} |e_i^\top e_j - s_{i,j}|. \tag{2}$$

Here the speaker embedding is L2 normalized, so the inner product is the cosine similarity.

4. Experiments

4.1. Experimental setup

We investigated cross-lingual TTS synthesis between English and Japanese. We first used LJ Speech corpus (English) [17] and JSUT corpus (Japanese) [18] for pretraining the cross-lingual TTS model, which was beneficial for model convergence. The speaker encoder and the rest of the model were trained separately in this stage. After the pretraining, we jointly trained all components using female speakers' speech included in VCTK [19] and JVS [20] corpora. We randomly chose 8 speakers (4 English and 4 Japanese) for unseen speaker evaluation and excluded them from the training set. The training set totally contained 107 speakers, in which 59 were English and 48 were Japanese. Different from the previous works, we used character as text representation, which was completely non-overlapping and language-dependent. All Chinese character in Japanese text was transformed to Hiragana by KyTea [21] to reduce character number in the data. To reduce computational cost, we downsampled all audios to 16 kHz. We also trained a WaveRNN neural vocoder [22] separately to convert 80-dimensional mel-spectrogram into time-domain waveform. The inter-speaker similarity score of JVS is included in the original dataset, the one of VCTK can be downloaded from internet.

In all our experiments, we used 64-dimensional speaker embedding, 16-dimensional language embedding. We normalized all speaker embeddings to unit length to stabilize the training process. The language classifier was a two-layer multi-layer perceptron. We found that simple architecture was more suitable for speaker classifier, so we used a linear classifier directly.

To condition Tacotron2 on speaker embedding and language embedding, we adopted the following modifications: (1) We concatenated language embedding with text embedding in the text encoder. (2) In the attention module, speaker embedding and language embedding were used as additional input. (3) In the decoder, we first used a linear transformation to transform the language embedding and context vector to a compact feature, then concatenated it with speaker embedding and fed it to the decoder. We found this could avoid possible entanglement of speaker embedding and language embedding.

We used Adam [23] as optimizer. During pretraining, we set the initial learning rate to 10^{-3} and got all embeddings from a randomly initialized embedding table. The batch size was set to 64. We multiplied the gradient back-propagated from language classifier to feature extractor by a factor $\lambda_p = \frac{2}{1+\exp(-10 \cdot p)} - 1$, where p represents the training progress ranging from 0 to 1. This allowed the feature extractor to be less influenced by the language classifier at the early training stage [14]. An L2 regularization term with 10^{-5} weight was added to the speaker encoder loss to avoid overfitting. After pretraining, we trained the model with 32 batch size on the whole training data and set the initial learning rate to 10^{-4}.

Four models are trained for comparison. Except for a baseline model trained by the conventional method ("Conv.") using speaker classifier only, two proposed models are trained using equation 1 and 2 separately, which are denoted by "DA" and "DA+sim" respectively. We finally train a model by adding similarity loss term to the baseline model, which is denoted by "Conv.+sim".

4.2. Subjective Evaluation

We evaluate naturalness and similarity score of the speech synthesized by each model. Each speaker has 20 intra-lingual utterances and 20 cross-lingual utterances for evaluation. All texts of these utterances are not included in the training data.

4.2.1. Naturalness Evaluation

We used five-level Mean Opinion Score (MOS) tests to evaluate the naturalness of speech. The MOS tests were conducted in each of all combinations of speaker settings (seen/unseen), languages (English/Japanese), and tasks (intra-/cross-lingual). 100 Japanese listeners participated in each evaluation, and 800 listeners participated in total. Each listener evaluated 25 utterances. The top half of Table 1 shows the result of intra-lingual TTS. We notice that our proposed models (DA+*) and Conv.+sim model obtained better scores than the baseline model (Conv.), though our method is not designed for intra-lingual synthesis. This improvement may be because the adversarial loss and similarity loss term have regularization effects on the model since they prevent speaker embedding distribute together and avoid possible overfitting.

Table 2: *Results of XAB tests on speaker similarity. The language column represents the language of synthesized speeches.* ***Bold*** *score indicates preferred method has p value less than* 0.05

Task	Language	Speaker	Conv. vs. DA	Conv.+sim vs. DA+sim	Conv. vs. Conv.+sim	DA vs. DA+sim
Intra-lingual	English	Seen	0.436 - **0.564**	0.512 - 0.488	0.488 - 0.512	**0.556** - 0.444
		Unseen	0.464 - 0.536	0.476 - 0.524	0.464 - 0.536	0.496 - 0.504
	Japanese	Seen	0.448 - **0.552**	**0.572** - 0.428	0.496 - 0.504	**0.580** - 0.420
		Unseen	0.416 - **0.584**	**0.608** - 0.392	0.360 - **0.640**	**0.568** - 0.432
Cross-lingual	English	Seen	0.420 - **0.580**	**0.588** - 0.412	0.440 - **0.560**	0.536 - 0.464
		Unseen	**0.584** - 0.416	**0.576** - 0.424	0.524 - 0.476	**0.568** - 0.432
	Japanese	Seen	**0.572** - 0.428	**0.552** - 0.448	0.524 - 0.476	0.508 - 0.492
		Unseen	0.468 - 0.532	0.484 - 0.516	0.432 - **0.568**	0.496 - 0.504

The bottom half of Table 1 shows the result of cross-lingual TTS. The score of ground truth score was relatively low, which may be because the listeners were all Japanese. Our proposed model obtained significant naturalness improvement when synthesizing cross-lingual speech. This demonstrates that our language-independent speaker embedding encodes more general voice information. We also notice that the DA+sim model sometimes fails to beat the baseline model. We consider this is because the similarity score between different languages doesn't exist, which introduces biases in the training process. All in all, our model can improve the naturalness of cross-lingual TTS synthesis, while maintaining the performance of intra-lingual TTS synthesis.

4.2.2. Speaker Similarity Evaluation

We conducted preference XAB tests to compare speaker similarity of the speech synthesized by each model. To study the effect of domain adaptation and perceptual similarity separately, we compared methods with and without domain adaptation, i.e., Conv. and DA, with and without speaker similarity regression, i.e., *+sim. 25 Japanese listeners participated in each test, and 800 listeners participated in total. Each listener evaluated 10 pairs of utterances. In all tests, we used the speech of the speaker's original language as ground truth.

The result is shown in Table 2. Our proposed model performs better than the baseline model in most intra-lingual cases and has comparable performance in cross-lingual cases when similarity loss is not used. This demonstrates that our proposed model has better speaker identity modeling ability and captures more general voice feature. However, we observe that when the similarity loss is added, our proposed model is beaten by the baseline model. A similar observation can be obtained from the rest two pairs, in which the baseline model is beaten by its similarity version, while our proposed model beats its similarity version. Again, we think this is because the lack of similarity score between different languages introduces more bias in the training process.

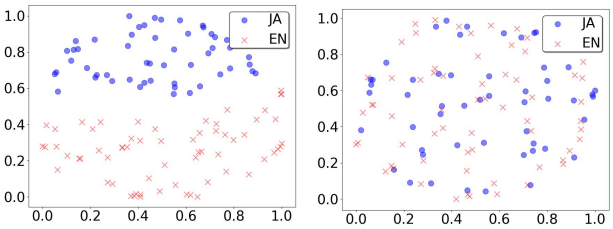
To sum up, the overall result demonstrates that our adversarial loss and perceptual similarity could improve the similarity score most of the time. However, how to combine these two methods to get more performance gains is unknown. We leave this as future work.

4.3. Speaker Space Evaluation

We evaluate speaker space generated by our method from various aspects. We first visualize English and Japanese speaker embedding with and without domain adaptation by the t-SNE algorithm [24]. The result is shown in Figure 4. We can see that, in the baseline model, the English speakers (circle) and Japanese speakers (cross) can be separated easily by languages. By using our method, speaker embeddings mix together and form a language-independent speaker space, which implies our method has better generalization ability when encountering multiple languages.

Table 3: *Results of XAB tests on speaker similarity of the nearest cross-lingual speaker pair.* ***Bold*** *score indicates preferred method has p value less than* 0.05

Conv. vs. DA	Conv. vs. DA+sim	DA vs. DA+sim
0.360 - **0.640**	0.256 - **0.744**	0.312 - **0.688**



(a) Conv. (b) DA

Figure 4: *t-SNE visualization of speaker embeddings learned by (a) baseline model and (b) proposed DA model.*

To verify whether our perceptual similarity regression contributes to transfer intra-lingual perceptual similarity knowledge to another language, we conducted preference XAB test to evaluate speaker similarity of the nearest cross-lingual speaker pair of each model. We assume that, if our method can construct a language-independent and human-interpretable speaker space, it is expected to find more perceptually similar speaker pairs beyond languages than the baseline model. We randomly picked 5 English speakers and found the Japanese speaker who has the highest cosine similarity with them in the speaker space generated by Conv., DA, and DA+sim models. Table 3 shows the result. Our DA model beats the baseline model, demonstrating that the speaker encoder trained by our domain adaptation method could capture more general voice information. Finally, our DA+sim model wins the DA model, which shows that the perceptual similarity could further improve the interpretability of the speaker space. All in all, our method could construct a language-independent and human-interpretable speaker space.

5. Conclusions

In this paper, we described cross-lingual TTS synthesis via domain adaptation and perceptual similarity regression in speaker space. Experimental results demonstrated that the speaker encoder of our model could not only improve synthesis performance but also find perceptually similar cross-lingual speaker pairs.

Acknowledgements: Part of this research and development work was supported by JSPS KAKENHI 18K18100 and 17H06101. JSPS and CAS under the Japan–People's Republic of China Research Cooperative Program.

6. References

[1] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 4052–4056.

[2] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NeurIPS*, Montréal, Canada, Dec. 2018, pp. 4480–4490.

[3] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Proc. NeurIPS*, Montréal, Canada, Dec. 2018, pp. 10 019–10 029.

[4] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 5901–5905.

[5] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting New Speakers Based on a Short Untranscribed Sample," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 3683–3691.

[6] B. Li and H. Zen, "Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis," in *Proc. Interspeech*, San Francisco, USA, Sept. 2016, pp. 2468–2472.

[7] H. Ming, Y. Lu, Z. Zhang, and M. Dong, "A light-weight method of building an LSTM-RNN-based bilingual TTS system," in *International Conference on Asian Language Processing (IALP)*, Singapore, Dec. 2017, pp. 201–205.

[8] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2080–2084.

[9] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 5621–5625.

[10] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-Lingual, Multi-Speaker Text-To-Speech Synthesis Using Neural Speaker Embedding," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 2105–2109.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Alberta, Canada, Apr. 2018, pp. 4779–4783.

[12] E. Nachmani and L. Wolf, "Unsupervised Polyglot Text-to-speech," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 7055–7059.

[13] S. Maiti, E. Marchi, and A. Conkie, "Generating Multilingual Voices Using Speaker Space Translation Based on Bilingual Speaker Data," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 7624–7628.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] Y. Saito, S. Takamichi, and H. Saruwatari, "DNN-based Speaker Embedding Using Subjective Inter-speaker Similarity for Multi-speaker Modeling in Speech Synthesis," in *Proc. 10th ISCA Speech Synthesis Workshop*, Vienna, Austria, Sep. 2019, pp. 51–56. [Online]. Available: \url{http://sython.org/demo/JSPS-DC1/index.html}

[16] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[17] K. Ito, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[18] 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.

[19] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[20] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv preprint arXiv:1908.06248*, 2019.

[21] G. Neubig and S. Mori, "Word-based partial annotation for efficient corpus construction," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[22] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 2410–2419.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.