Exploring Cross-lingual Singing Voice Synthesis Using Speech Data

Yuewen Cao[1,*], *Songxiang Liu*[1,*], *Shiyin Kang*[3], *Na Hu*[2], *Peng Liu*[2]
Xunying Liu[1], *Dan Su*[2], *Dong Yu*[2], *Helen Meng*[1]

[1] The Chinese University of Hong Kong, [2]Tencent AI Lab, [3]Huya Inc.

{ywcao,sxliu,xyliu,hmmeng}@se.cuhk.edu.hk, kangshiyin@huya.com,
{ninahu,feanorliu,dansu,dyu}@tencent.com

Abstract

State-of-the-art singing voice synthesis (SVS) models can generate natural singing voice of a target speaker, given his/her speaking/singing data in the same language. However, there may be challenging conditions where only speech data in a non-target language of the target speaker is available. In this paper, we present a cross-lingual SVS system that can synthesize an English speaker's singing voice in Mandarin from musical scores with only her speech data in English. The presented cross-lingual SVS system contains four parts: a BLSTM based duration model, a pitch model, a cross-lingual acoustic model and a neural vocoder. The acoustic model employs encoder-decoder architecture conditioned on pitch, phoneme duration, speaker information and language information. An adversarially-trained speaker classifier is employed to discourage the text encodings from capturing speaker information. Objective evaluation and subjective listening tests demonstrate that the proposed cross-lingual SVS system can generate singing voice with decent naturalness and fair speaker similarity. We also find that adding singing data or multi-speaker monolingual speech data further improves generalization on pronunciation and pitch accuracy.

Index Terms: singing voice synthesis, cross-lingual, encoder-decoder, adversarial loss

1. Introduction

The goal of singing voice synthesis (SVS) is to generate singing voice from musical scores with lyrics. In recent years, singing synthesis technologies have made rapid progress with successful application of deep learning techniques [1, 2, 3, 4, 5, 6, 7], generating synthesized songs with high quality. Researchers have extended these SVS systems to enable control of speaker identity [8, 9, 10] and singing style [11, 12]. However, extending such models to support cross-lingual singing synthesis with a target voice is non-trivial. There may be challenging cases where only the speech data in a non-target language of the target speaker is available, i.e., synthesizing an English speaker's singing voice in Mandarin from musical scores with only his/her speech data in English. This occurs when, for example, the target speaker is unable to sing properly, or he/she cannot speak in the target language.

Most existing SVS systems only support one language. There have been several promising results in using encoder-decoder based monolingual SVS to synthesize the target speaker's singing voice given only his/her speech samples in the same language [13, 14]. In [13], the Tacotron2 GST model is extended with speaker embedding and pitch contours for singing synthesis, where only speech data is used during training. The tasks of speech synthesis and singing synthesis are intergrated into a unified framework with learned shareable speaker embeddings between speech and singing synthesis [14].

There are few examples of cross-lingual SVS systems using target speaker's speech data in the literature. In [15], a bilingual Japanese and English SVS system is built with the hidden Markov model (HMM) using singing data from a bilingual singer. However, in practice, it is hard and expensive to obtain such bilingual singing data in large quantities. A recent proposed multi-lingual multi-singer SVS system is built with multi-singer Mandarin, English and Cantonese singing data mined from music websites [16]. The system contains several steps, including data crawling, singing and accompaniment, lyrics-to-singing alignment, data filtration and singing modeling. As a byproduct of multi-lingual training, it can perform cross-lingual SVS for a designated singer. Instead of synthesizing from music scores, the system needs demo singing audio to extract the pitch and phoneme duration information during inference. Also, the ability of cross-lingual SVS using a target speaker's voice has not been studied in [16]. Compared with speech data, singing data is much more difficult and costly to collect. With easy access to existing large-scale high-quality monolingual speech corpora, we intend to investigate the use of monolingual speech corpora from different speakers for this cross-lingual SVS task.

SVS bears similarity to text-to-speech (TTS) synthesis in terms of producing natural voice from textual information input, but SVS has the musical note to constrain the pitch and duration of each syllable. Both SVS and TTS face the problem of cross-lingual synthesis, where bilingual or multilingual data is not available. Recent progress of encoder-decoder architectures has achieved a resounding success in cross-lingual TTS on mix of monolingual corpora [17, 18, 19]. The decoders in these methods are conditioned on a speaker embedding to control speech voice, while the encoders employ different mechanisms to handle different language inputs. It is found in [19] that phoneme-based model performs better in rare words and out-of-vocabulary (OOV) situations than byte and character counterparts. A speaker-adversarial loss term is used to encourage the model to disentangle speaker identity representation from the text content. This model can consistently synthesize intelligible and native speech for training speakers in all languages seen during training.

In this paper, we explore cross-lingual SVS using the target speaker's speech data based on the encoder-decoder architecture. Specifically, given an English speaker's English speech data, our objective is to build an SVS system which can synthesize the English speaker's singing voice in Mandarin from musical scores. The proposed cross-lingual SVS system contains four parts: a bidirectional long-short term memory (BLSTM) based duration model, a pitch model, an encoder-decoder based cross-lingual multi-speaker acoustic model and a neural vocoder. The BLSTM based duration model predicts

[*] Work done during internship at Tencent AI Lab

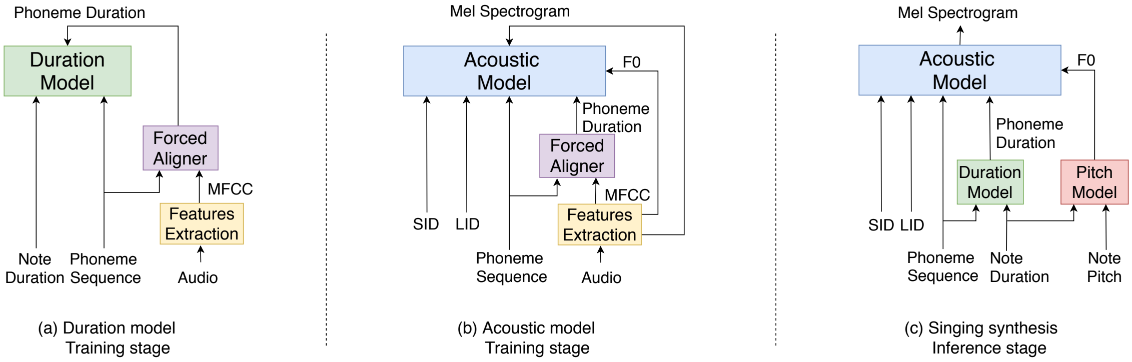

Figure 1: *The training stage of duration model and acoustic model, and the inference stage of the proposed cross-lingual singing voice synthesis system.*

phoneme duration from the phoneme sequences and note duration. The singing pitch model transforms the musical note to the fundamental frequency (F0). The cross-lingual multi-speaker acoustic model is based on the cross-lingual TTS model with speaker adversarial loss [19]. The acoustic model generates mel-spectrograms from a phoneme sequence, conditioned on the F0, phoneme duration, speaker embedding and language embedding. The neural vocoder finally converts the mel-spectrograms into time-domain waveforms. We also investigate the effectiveness of including additional singing data or multi-speaker monolingual data in acoustic model training on cross-lingual SVS performance.

The rest of the paper is organized as follows: Section 2 introduces the detailed structure of proposed cross-lingual SVS system. Section 3 describes experimental details and the evaluation results. Conclusions are drawn in section 4.

2. Proposed system

The proposed cross-lingual SVS system is composed of four parts: a duration model, a pitch model, a cross-lingual multi-speaker acoustic model and a neural vocoder, which are introduced in the following subsections. The training stages of duration model and acoustic model, and the inference stage of the proposed system are illustrated in Figure 1.

2.1. Duration model

In singing synthesis, the phoneme duration is strongly constrained by the musical notes, which is a notable difference from TTS. The duration model takes phoneme sequence and note duration as inputs to predict the singing duration of each phoneme. The note duration is converted into frame count (the number of frames of each syllable). The phoneme sequence is concatenated with the corresponding syllable frame count, and then sent to the duration model as inputs. The duration model consists of one fully connected (FC) layer with ReLU activation and dropout, followed by two BLSTM layers. The training stage of duration model is shown in Figure 1 (a). Audio and its corresponding phoneme sequence are aligned by a forced aligner. The duration of each phoneme is measured by the number of aligned frames. We minimize the mean squared error (MSE) between the predicted phoneme duration and the duration obtained from the forced aligner to train the duration model. During inference, an additional post-processing step is performed after the duration model to ensure that the sum of predicted phoneme duration matches the target note duration as in [1].

2.2. Pitch model

Pitch is among the most important perceptual components of singing voices, whose variation is associated with musical melodies. Additionally, phonetics can also cause inflection in pitch contours, so-called microprosody [20]. In this study, we simply model pitch from music notes with some heuristic rules. We convert the note pitch to F0 and expand it to frame-level according to the corresponding note duration. Then we convolve the F0 sequence with a triangular window aligned with the centering frame.

2.3. Acoustic model

As shown in Figure 2, we adopt an encoder-decoder based acoustic model with speaker classifier to generate mel-spectrograms from input phoneme sequences, conditioned on a speaker embedding, a language embedding, F0 and phoneme duration. The text encoder takes phoneme sequences as input and adopts CBHG architecture as in [21]. The CBHG module consists of a bank of convolutional filters, highway networks and a bidirectional gated recurrent unit (GRU). Following [19], an adversarially-trained speaker classifier is employed to discourage the text encodings from capturing speaker information. The speaker classifier is optimized with the objective: $\mathcal{L}_{\text{speaker}}(\psi_S; \mathbf{t}_i) = \sum_{i=1}^{N} \log p(\mathbf{s}_i|\mathbf{t}_i)$, where ψ_S are the parameters of the speaker classifier, $\boldsymbol{s_i}$ is the speaker label corresponding to encoder outputs $\boldsymbol{t_i}$ and N is the number of training samples. To jointly train the speaker classifier and remaining parts of the acoustic model, a gradient reversal layer is added prior to the speaker classifier, which scales the gradient by $-\lambda$. Though it is suggested that adding a residual encoder improves stability and naturalness of cross-lingual transfer in [19], our preliminary experiments show that the residual encoder does not bring improvement. We omit the residual encoder in our model. To ensure the hard alignments between the phoneme sequence, musical note duration and the corresponding acoustic features, the decoder is explicitly conditioned on phoneme duration with the attention part omitted in our system. The text encodings are expanded by replicating hidden states sequentially along time axis according to the phoneme duration as in [22]. The frame-level F0 goes through a FC layer with ReLU activation and dropout before being sent to the decoder.

The decoder is an autoregressive recurrent neural network (RNN), which is composed of a pre-net layer, two LSTM decoder layers and output layers following [23]. The prediction

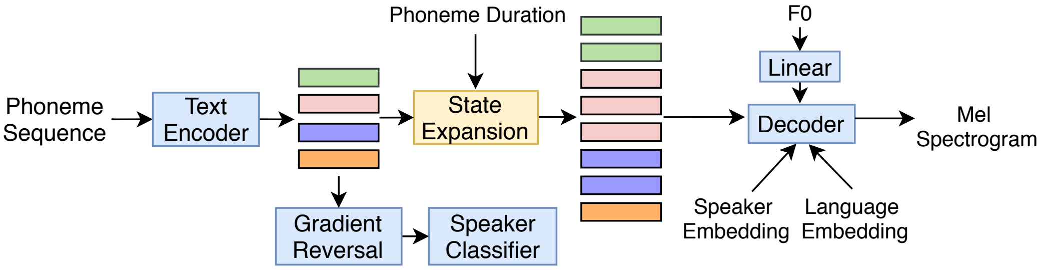

Figure 2: *Architecture of the acoustic model.*

from the previous time step is first passed to the pre-net as input of current time step. The text encodings and F0 of current time step are concatenated with the pre-net output, speaker embedding and language embedding as decoder input, which is sent to the LSTM decoder layers. Then, the concatenation of the LSTM output, the text encodings and F0 of current time step is projected through a linear transformation to predict the target spectrogram frame. The predicted features are passed through a convolutional post-net to predict a residual to add to the initial prediction. The adversarial loss from the speaker classifier and MSE losses from before and after the post-net are summed up to optimize the whole acoustic model. The speaker embedding lookup table and language embedding lookup table are jointly learned with the acoustic model. As shown in Figure 1 (b) and (c), the phoneme duration and F0 input to the acoustic model are extracted from audio during the training stage, while predicted from singing duration model and pitch model respectively during the inference stage.

3. Experiments

3.1. Corpora

In this paper, our goal is to study how to build an SVS system that can synthesize the English speaker's singing voice in Mandarin from musical scores, given only her English speech data. Three corpora including the multi-speaker English corpus VCTK [24], an internal multi-speaker Mandarin speech corpus and a Mandarin singing corpus are involved in our experiments. The VCTK corpus contains 44 hours of clean speech from 109 English speakers. Each English speaker has a varied number of utterances from 200 to 500. We choose a female speaker p261 from VCTK as the target speaker. The internal Mandarin speech corpus contains around 46 hours of clean speech from 82 Mandarin speakers. Each Mandarin speaker has around 500 utterances. The internal Mandarin singing corpus is recorded by a female singer, which contains 9 hours of exercise songs. The exercise songs are short utterances, which have good coverage for phoneme, pitch and note duration.

3.2. Experimental setup

The musical scores containing event tuples with pitch, note duration and syllables, which are described in MusicXML format [25]. The text inputs and lyrics are transcribed into phoneme sequences through text analysis procedures. Audio is sampled at 24 kHz with leading and trailing silence trimmed. The 80-band mel-spectrograms are extracted with 50ms window shifted by 10ms. We use continuous F0 with frame shift of 10ms. The phoneme duration is obtained by an internal HMM-GMM based forced aligner from the 39-dimensional MFCCs.

For duration model, one FC layer activated by ReLU with dropout 0.5 containing 512 units and two BLSTM layers with 256 hidden units in each direction are used. The duration model is trained with the Mandarin singing corpus. For acoustic model, the text encoder strictly follows the encoder architecture in [21]. The speaker classifier has one FC layer with 256 units activated by ReLU, followed by another FC layer activated by softmax. The output dimension of the final FC layer is the number of training speakers. The loss weight of speaker classifier and the gradient scale factor are set to 0.02 and 0.5 respectively. The dimensions of speaker embedding and language embedding are set to 64 and 2 respectively. F0 passes through a 128-unit FC layer with dropout 0.5 activated by ReLU before being sent to the decoder. The decoder is implemented as [23]. We use a WaveRNN [26] vocoder to synthesize waveforms from the predicted mel-spectrograms. The WaveRNN vocoder is trained using all the training data. The duration model, acoustic model and neural vocoder are trained separately.

We aim to investigate the proposed cross-lingual SVS system from two aspects: 1) whether the system can synthesize natural target speaker's singing voice in the target language; 2) whether the acoustic model gains from including singing data or multi-speaker speech data in acoustic model training. We implement our proposed systems using different corpora, thus bringing the following four systems:

- $System_1$: The acoustic model is trained with English speech data from speaker p261 and Mandarin singing data from one singer.
- $System_2$: The acoustic model is trained with English speech data from 109 speakers including speaker p261 and Mandarin speech data from 82 speakers.
- $System_3$: The acoustic model is trained with English speech data from 109 speakers including speaker p261, Mandarin speech data from 82 speaker and Mandarin singing data from one singer.
- $Ablation$: The acoustic model has no speaker classifier and is trained with English speech data from 109 speakers including speaker p261 and Mandarin speech data from 82 speaker.

3.3. Evaluation and analysis

We evaluate how well the proposed cross-lingual SVS system can be used to synthesize songs in Mandarin with the English speaker p261's voice. Three Mandarin pop songs are synthesized using each system with speaker embedding of speaker p261. The input F0s to the systems are adjusted by two keys lower to account for the speaker's vocal range. Objective and subjective evaluations on all four systems are conducted.

Table 1: *Root-mean-square error (RMSE) and Pearson correlation (CORR) coefficient results between F0 extracted from synthesized songs and the F0 input to acoustic model in* $System_1$, $System_2$ *and* $System_3$.

	$System_1$	$System_2$	$System_3$	$Ablation$
F0 RMSE (Hz)	25.34	16.797	**14.382**	16.475
F0 CORR	0.945	0.975	**0.981**	0.976

3.3.1. Objective evaluation

We evaluate whether the acoustic models in the four systems can produce accurate singing voice for a given input. We extract F0 sequence from the generated songs and compare it to the acoustic model input F0. The more similar are two sequences, the more precisely the acoustic model generates singing voice conditioned on the input F0. Due to duration conditioning, the extracted F0 sequence and input F0 sequence have the same length and do not require padding. Table 1 lists the root-mean-square error (RMSE) and the Pearson correlation (CORR) coefficient between the F0 extracted from synthesized songs and the F0 input to acoustic model. The high F0 correlation coefficients of all systems show that our proposed cross-lingual acoustic model can generate singing voice with precise pitch and timing for the given input. $System_3$ achieves smaller pitch distortion and higher correlation coefficient than $System_2$ and $System_1$. This indicates that including singing data and multi-speaker speech data benefits acoustic model training. $System_2$ and $Ablation$ have comparable pitch distortion and correlation coefficient, which indicates that the speaker classifier brings little effect on pitch accuracy.

3.3.2. Subjective evaluation

Two mean opinion score (MOS) tests are conducted for subjective evaluation of naturalness and speaker similarity of the synthesized songs with the target speaker's voice. The synthesized songs are segmented into short utterances for the convenience of evaluation, from which 20 audio samples are randomly selected for testing, ranging from 5s to 10s. 15 native Mandarin speakers participated in the subjective listening tests. All generated songs are synthesized with the predicted phoneme duration and F0 ("http://demo-page.github.io/crosslingualSVS").

Naturalness. In the MOS test, the subjects listen to each pair of 4 audio samples synthesized by the four systems and give a 5-point scale score of naturalness (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The lyrics of audio samples are provided for more accurate pronunciation evaluation. As shown in Figure 3, both $System_2$ and $System_3$ can significantly improve the naturalness of synthesized songs than $System_1$. This indicates the acoustic model benefits a lot from additional multi-speaker English and Mandarin speech data. According to listeners' comments, most of the degradation comes from the unclear pronunciation, not audio fidelity or accent. Listeners comment that some audio samples sound like humming. This validates that increasing training speaker diversity improves cross-lingual pronunciation generalization. Figure 3 also shows that $System_3$ has slight improvement over $System_2$ in naturalness. Listeners comment that the improvement comes mostly from more natural high tones. Singing data contains a much higher range of pitches, and benefits the acoustic model in pitch generalization across speakers. $System_2$ and $Ablation$ have similar naturalness.

Speaker similarity. Another 5-point MOS test is conducted similar to the one described above. Listeners are instructed to ignore the content and focus on the speaker similarity between the reference utterances and presented audio samples. Note that this similarity evaluation is more challenging than the one in [19], due to the significant difference between singing and speaking voices for most people [27, 28]. We provide the ground truth English speech utterances and synthesized Mandarin speech utterances of speaker p261 as reference utterances. Results are presented in Figure 4. Both $System_2$ and $System_3$ can achieve fair speaker similarity, albeit with significantly reduced performance of $System_1$ and $Ablation$. This validates that increasing training speaker diversity consistently improves cross-lingual generalization in terms of naturalness and speaker similarity. Adding a speaker classifier can improve speaker similarity.

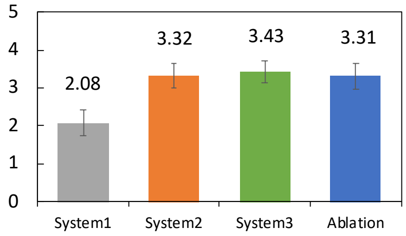

Figure 3: *MOS results with 95% confidence intervals on naturalness for* $System_1$, $System_2$, $System_3$ *and* $Ablation$*. Higher is better.*

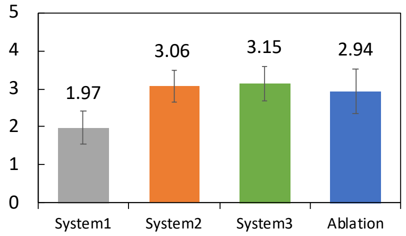

Figure 4: *MOS results with 95% confidence intervals on speaker similarity for* $System_1$, $System_2$, $System_3$ *and* $Ablation$*. Higher is better.*

4. Conclusions

In this paper, we present a cross-lingual SVS system that can generate songs in Mandarin with an English speaker's voice. Only English speech data from the target speaker is available for model training. The proposed cross-lingual SVS system contains four parts: a BLSTM based duration model, a pitch model, an encoder-decoder based cross-lingual multi-speaker acoustic model and a neural vocoder. Objective and subjective experimental results validate the effectiveness of the proposed system in terms of naturalness and speaker similarity. Adding multi-speaker data or singing data can further improve the generalization on pronunciation and pitch accuracy. In the future, we will explore more techniques for the cross-lingual voice timbre retaining. Techniques for pitch prediction from musical note to cover more singing styles will also be investigated.

5. Acknowledgements

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14207619).

6. References

[1] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Applied Sciences*, 2017.

[2] Y. Yi, Y. Ai, Z. Ling, and L. Dai, “Singing voice synthesis using deep autoregressive neural networks for acoustic modeling,” *Interspeech*, 2019.

[3] J. Lee, H. Choi, C. Jeon, J. Koo, and K. Lee, “Adversarially trained end-to-end korean singing voice synthesis system,” *Interspeech*, 2019.

[4] K. Nakamura, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “Fast and high-quality singing voice synthesis system based on convolutional neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] M. Blaauw and J. Bonada, “Sequence-to-sequence singing synthesis using the feed-forward transformer,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, “Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders,” *arXiv preprint arXiv:2004.11012*, 2020.

[7] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, “Xiaoicesing: A high-quality and integrated singing voice synthesis system,” *arXiv preprint arXiv:2006.06261*, 2020.

[8] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, “Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan,” in *European Signal Processing Conference (EUSIPCO)*, 2019.

[9] M. Blaauw, J. Bonada, and R. Daido, “Data efficient voice cloning for neural singing synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[10] J. Wu and J. Luan, “Adversarially trained multi-singer sequence-to-sequence singing synthesizer,” *arXiv preprint arXiv:2006.10317*, 2020.

[11] J. Lee, H.-S. Choi, J. Koo, and K. Lee, “Disentangling timbre and singing style with multi-singer singing synthesis system,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[12] J. Bonada and M. Blaauw, “Hybrid neural-parametric f0 model for singing synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[13] R. Valle, J. Li, R. Prenger, and B. Catanzaro, “Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,” in *IEEE International Conference on Acoustics, Speech and Signal, Processing (ICASSP)*, 2020.

[14] L. Zhang, C. Yu, H. Lu, C. Weng, Y. Wu, X. Xie, Z. Li, and D. Yu, “Learning singing from speech,” *arXiv preprint arXiv:1912.10128*, 2019.

[15] K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda, “Hmm-based singing voice synthesis and its application to japanese and english,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[16] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, “Deepsinger: Singing voice synthesis with data mined from the web,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1979–1989.

[17] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[18] E. Nachmani and L. Wolf, “Unsupervised polyglot text-to-speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[19] 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,” *Interspeech*, 2019.

[20] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[21] Y. Wang, R. Skerry Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, “Tacotron: Towards end-to-end speech synthesis,” *Interspeech*, 2017.

[22] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, “Durian: Duration informed attention network for multimodal synthesis,” *arXiv preprint arXiv:1909.01700*, 2019.

[23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[24] V. Christophe, Y. Junichi, and M. Kirsten, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” *The Centre for Speech Technology Research (CSTR)*, 2016.

[25] “Musicxml definition,” https://www.musicxml.com/.

[26] 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 *International Conference on Machine Learning*, 2018.

[27] S. Rosenau, “An analysis of phonetic differences between german singing and speaking voices,” in *Int. Congress of Phonetic Sciences*, 1999.

[28] W. H. Tsai and H. C. Lee, “Singer identification based on spoken data in voice characterization,” *IEEE transactions on audio, speech, and language processing*, 2012.