CODE-SWITCHED SPEECH SYNTHESIS USING BILINGUAL PHONETIC POSTERIORGRAM WITH ONLY MONOLINGUAL CORPORA

Yuewen Cao[*1], *Songxiang Liu*[1], *Xixin Wu*[1], *Shiyin Kang*[3], *Peng Liu*[3]
Zhiyong Wu[†1,2], *Xunying Liu*[1], *Dan Su*[3], *Dong Yu*[3], *Helen Meng*[1]

[1]Human-Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[3]Tencent AI Lab, Tencent, Shenzhen, China
{ywcao,sxliu,wuxx,zywu,xyliu,hmmeng}@se.cuhk.edu.hk,{shiyinkang,feanorliu,dansu,dyu}@tencent.com

ABSTRACT

Synthesizing fluent code-switched (CS) speech with consistent voice using only monolingual corpora is still a challenging task, since language alternation seldom occurs during training and the speaker identity is directly correlated with language. In this paper, we present a bilingual phonetic posteriorgram (PPG) based CS speech synthesizer using only monolingual corpora. The bilingual PPG is used to bridge across speakers and languages, which is formed by stacking two monolingual PPGs extracted from two monolingual speaker-independent speech recognition systems. It is assumed that bilingual PPG can represent the articulation of speech sounds speaker-independently and captures accurate phonetic information of both languages in the same feature space. The proposed model first extracts bilingual PPGs from training data. Then an encoder-decoder based model is used to learn the relationship between input text and bilingual PPGs, and the bilingual PPGs are mapped to acoustic features using bidirectional long-short term memory based model conditioned on speaker embedding to control speaker identity. Experiments validate the effectiveness of the proposed model in terms of speech intelligibility, audio fidelity and speaker consistency of the generated code-switched speech.

Index Terms— code-switching, speech synthesis, phonetic posteriorgrams

1. INTRODUCTION

Code-switching (CS), the alternation of languages within an utterance, is a common phenomenon in multilingual societies across the world [1]. State-of-the-art text-to-speech (TTS) synthesis models can generate monolingual speech with high intelligibility and naturalness [2–5]. However, they are not fully capable of handling code-switched text, which results in omitted or incorrect pronunciations in the synthesized outputs.

It is straightforward to use bilingual recordings from a bilingual speaker for building a code-switched speech synthesizer [6–9]. However, in practice, it is expensive to obtain such bilingual data in large quantities. With easy access to large-scale existing monolingual corpora, we intend to investigate the combined use of monolingual recordings from different speakers to generate code-switched speech. In this setting, speaker characteristics and language characteristics are directly correlated, i.e., each speaker speaks in only one language. This makes it difficult to transfer voices when there is language switching, which can easily lead to synthesize outputs with inconsistent voices within an utterance. Another challenge is the mismatch between training and testing. The model is trained with monolingual corpora of the code-switched language pairs, and never accesses language alternations during training. Such a model will often fail to generate intelligible speech with smooth transitions at language boundaries during testing.

One possible solution to tackle the aforementioned issues in CS TTS is to find proper speaker-independent phonetic features which are able to represent the code-switched languages in a compact feature space. Phonetic PosteriorGrams (PPGs) computed from a speaker-independent automatic speech recognition (SI-ASR) model are deemed speaker-independent and language-independent, which have the potential to serve as such a representation. The PPG is a time-versus-class matrix representing the posterior probabilities of each phonetic class for a specific frame of one utterance [10, 11]. It is considered to be speaker-independent since the SI-ASR model is designed to be invariant with different speakers. Benefiting from its speaker-independent property, PPGs have been successfully used in voice conversion (VC) tasks [12–14]. [8, 15–17] indicates that state-level and frame-level speech segments can be shared in different languages. Hence PPGs, which are frame-level features, can be regarded as language-independent. This property of PPGs has made it possible to be used in the cross-lingual VC task [18]. However, languages are phonetically different in nature, meaning that PPGs of one language cannot effectively characterize the phonetic contents of another language. A recent cross-lingual VC system uses bilingual PPGs as the linguistic representations to accurately capture the phonetic information of both languages [19]. The bilingual PPGs are formed by stacking two monolingual PPGs extracted from two monolingual SI-ASR systems. Bilingual PPGs can offer a representation superior to monolingual PPGs for representing different languages in a speaker-independent and language-independent space.

In this paper, we explore the efficacy of using bilingual PPGs

[*]Work done during internship at Tencent
[†]Corresponding author

for code-switched TTS using a combination of Mandarin and English monolingual speech corpora uttered by two female speakers. Specifically, the attention-based encoder-decoder model Tacotron2 [4] is adapted to convert input text sequences to bilingual (Mandarin and English) PPGs. The bilingual PPGs are then mapped to acoustic features frame-wise using a bidirectional long-short term memory (BLSTM) based model, conditioned on speaker embedding to control the speaker identity. Bilingual PPGs are inherently speaker and language disentangled, making it easy to control speaker consistency for code-switched speech. Besides, the textual content of an utterance is closely related to its bilingual PPGs, making it much easier to learn the mapping from text to bilingual PPGs than the mapping from text to acoustic features, using an encoder-decoder model. This is helpful for accelerating model training and reducing sequence-to-sequence alignment errors during code-switched speech generation. Experiments validate the effectiveness of the proposed model in terms of speech intelligibility, audio fidelity and speaker consistency of the generated code-switched speech.

2. RELATED WORK

Early attempts mostly adopt HMM-based and unit-selection-based TTS models, including voice adaptation from an average voice [20], voice conversion to create a polyglot corpus from monolingual corpora [21], and unit mapping methods, e.g. phoneme mapping [22], frame mapping [23], senone mapping [17], etc. We borrow the idea from these works in equalizing the speaker differences. Recent approaches take advantage of the encoder-decoder architecture for code-switched and cross-lingual TTS without training on bilingual or multilingual data. The decoders in these methods are conditioned on a speaker embedding to control speech voice with mel spectrograms as output, while the encoders employ different mechanisms to handle different language inputs. [24] uses Unicode byte representation for all languages. Separate encoder and shared encoder with language embedding are investigated in [25] to encode individual languages with characters as input. [26, 27] use phoneme sequences as input, while [26] starts from an average voice model built from multi-speaker monolingual data and [27] incorporates latent variables into the attention mechanism to generate language agnostic articulatory features to improve generalization during inference. Since mel spectrograms inherently capture both spoken language characteristics and speaker identity, these encoder-decoder models with mel spectrograms as output implicitly disentangle language characteristics and speaker characteristics. To explicitly encourage the model to learn disentangled representation of the text and speaker identity, an adversial loss is used in [28]. Our work is similar to the adversial loss based model in terms of explicit speaker and language disentanglement.

3. BASELINE APPROACH

The baseline approach follows a Tacotron2-based cross-lingual voice cloning model [28], where a speaker-adversial loss is incorporated to disentangle closely correlated textual content and speaker identity. Phoneme sequences are used as input, since [28] has shown that phoneme-based model performs better in rare words and out of vocabulary (OOV) situations than byte and character counterparts.

As shown in Fig. 1, a text encoder takes phoneme sequences as input. The text encodings are sent to an adversarially-trained speaker classifier for discouraging the text encodings from capturing speaker information. The speaker classifier is optimized with the objective:

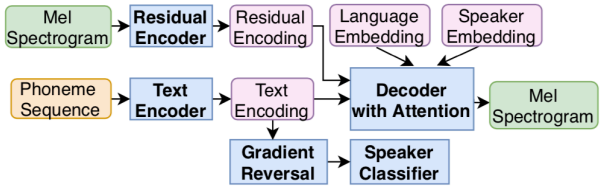

Fig. 1. The baseline CS TTS model

$\mathcal{L}_{\text{speaker}}(\psi_S; \mathbf{t}_i) = \sum_i^N \log p(\mathbf{s}_i|\mathbf{t}_i)$, where ψ_S are the parameters of the speaker classifier, $\boldsymbol{s_i}$ is the speaker label corresponding to text encoding $\boldsymbol{t_i}$ and N is the number of training samples. Prior to the speaker classifier, we add a gradient reversal layer, which scales the gradient by $-\lambda$. The text encodings are also accessed by the decoder through a location-sensitive attention [29], which takes attention history into account when computing the attention probabilities for aggregation. The residual encodings, which are encoded from mel spectrograms by a variational autoencoder (VAE)-like residual encoder [30], are used to help stabilize attention. The aggregated text encodings, residual encodings, language embedding and speaker embedding are concatenated at each decoder time step. The decoder takes these as input to generate mel spectrograms autoregressively, and also predicts an end-of-sentence flag at each time step. The spectrograms and stop flags are trained with mean squared error (MSE) loss L_{mel} and binary cross entropy (BCE) loss L_{stop} respectively. The baseline model is jointly trained using the loss L_{TTMEL}:

$$L_{TTMEL} = \alpha_1 L_{mel} + \alpha_2 L_{stop} + \alpha_3 L_{vae-KL} + \alpha_4 L_{speaker}, \quad (1)$$

where L_{vae-KL} is the Kullback-Leibler divergence loss in VAE training and the αs are weights of the four losses.

4. CODE-SWITCHED TTS WITH BILINGUAL PPG

We introduce bilingual PPGs for code-switched TTS, leveraging its speaker-independent and language-independent properties. As illustrated in Fig. 2, the proposed approach includes three parts: bilingual PPG extraction with two monolingual SI-ASR models, an attention-based encoder-decoder model mapping text to bilingual PPGs, and a BLSTM-based model mapping bilingual PPGs to mel spectrograms.

4.1. Bilingual PPG Extraction

Fig. 2(a) presents the procedure of bilingual PPG extraction. Two DNN-HMM based SI-ASR models (English and Mandarin) are pretrained by an English ASR corpus and a Mandarin ASR corpus respectively. Monolingual PPGs are first extracted with the English and Mandarin SI-ASR models separately. Then the two monolingual PPGs are stacked to form a bilingual PPG, which represents speaker-independent articulation of speech sounds from both languages in a compact space.

4.2. Text-to-Bilingual PPG

We modify the encoder-decoder model Tacotron2 to generate bilingual PPGs from input phoneme sequences. As shown in Fig. 2(b), the decoder predicts a bilingual PPG, a log F0 (LF0), a voice/unvoiced (VUV) flag and an end-of-sentence flag from the encoded phoneme sequence one frame at a time. The attention-based decoder is composed of a pre-net layer, a location-sensitive attention layer, two LSTM decoder layers and output layers, following [4]. The prediction of bilingual PPG from the previous time step is

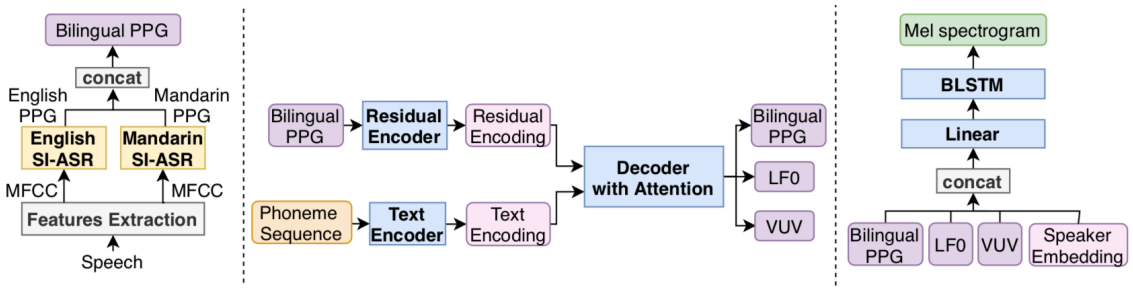

Fig. 2. The proposed bilingual PPG based CS TTS model. (a) The process of computing bilingual PPGs from the speech signal. (b) Encoder-decoder based model mapping text inputs to bilingual PPGs. (c) BLSTM-based model mapping bilingual PPGs to mel spectrograms.

passed to the pre-net as input of current time step. The pre-net output and attention context are concatenated as decoder input, which is sent to the LSTM decoder layers. Then, the concatenation of LSTM output and attention contexts is projected through two separate linear transformations to predict the target bilingual PPGs and LF0 respectively. Meanwhile, the concatenation is also projected into two scalars by the sigmoid activation to predict an end-of-sentence flag and a VUV flag respectively. LF0 and VUV are also predicted here to compensate for the prosody information lacking in bilingual PPGs. MSE loss is adopted for bilingual PPG (L_{bppg}) and LF0 (L_{lf0}) predictions, while BCE loss is used for VUV flag (L_{vuv}) and stop flag (L_{stop}) predictions. A VAE-like residual encoder is also adopted for stabilizing attention, which encodes the latent factors from bilingual PPGs. The residual encoder closely follows the network architecture in [30]. The text-to-bilingual PPG model is trained using the loss L_{TTBPPG}:

$$L_{TTBPPG} = \alpha_1 L_{bppg} + \alpha_2 L_{lf0} + \alpha_3 L_{vuv} + \alpha_4 L_{stop} + \alpha_5 L_{vae-KL}, \quad (2)$$

where L_{vae-KL} is the Kullback-Leibler divergence loss in VAE training and the αs are weights of the five losses.

4.3. Bilingual PPGs to Mel spectrograms

A BLSTM based transformation model is used to map bilingual PPGs to mel spectrograms, as shown in Fig. 2(c). A speaker embedding is concatenated with bilingual PPG, LF0 and VUV at each frame to control the speech voice, before being sent to the transformation model. LF0 and VUV compensate for the prosody information lacking in bilingual PPGs. Following [13], the transformation model comprises of two fully connected (FC) layers with ReLU activation and dropout, followed by four BLSTM layers. The speaker embedding is jointly trained with the transformation model. During synthesis, the predicted bilingual PPG, LF0 and VUV from the text to bilingual PPG model are concatenated with the designated speaker embedding to control the speaker identity of the generated speech.

5. EXPERIMENTS

5.1. Experimental setup

We use an American English speech corpus [31] and a Mandarin speech corpus [32] uttered by two female speakers to build our systems. All audios are sampled at 16 kHz with leading and trailing silence trimmed. 9000 utterances from each corpus are randomly selected as training data, and both corpora have about 10 hours of speech. Another 300 utterances from each corpus and 300 code-switched utterances crawled from the Internet are used as test data. Each CS utterance has one or two code-switched points. English words and Chinese characters are transcribed as phonemes as input with stress and tonal information respectively. We extract F0 and VUV flag with frame shift of 10ms. Then F0 is linearly interpolated and transformed to logarithmic scale before being normalized to have zero mean and unit variance over each corpus. The 80-band mel spectrograms are extracted with 25ms window shifted by 10ms.

We implement the baseline model following [28]: the weights in Equation (1) are set to 1.0, 1.0, 1.0 and 0.02 respectively, and the gradient scale factor is set to 0.5. The only modifications are the dimensions of speaker embedding and language embedding, which are set to 16 and 2 respectively, since we only have two speakers and two languages. We use the open-source WaveRNN network[1] as neural vocoder to invert mel spectrograms to waveforms. Two separately trained WaveRNNs with ground-truth mel spectrograms for each speaker are used for both the baseline and proposed models.

The English and Mandarin ASR models are trained on TIMIT [33] and AI-SHELL1 [34] corpus respectively. Both SI-ASR models are implemented with the Kaldi toolkit [35]. 13-dimensional MFCCs computed with 25ms window shifted by 10ms are used for both SI-ASR models, while Mandarin SI-ASR also takes 3 dimensional pitch features as input following [34]. The English SI-ASR model has 4 hidden layers with 1024 hidden units, while the Mandarin SI-ASR model consists of 6 hidden layers with 850 hidden units. Senones are treated as the phonetic class of PPGs. The number of senone classes for English and Mandarin ASR is 128 and 217 respectively, which are both obtained by clustering at the SI-ASR training stage. The bilingual PPGs, formed by concatenating English PPGs and Mandarin PPGs, thus have 345 dimensions. The text-to-bilingual PPG model closely follows the network architecture of Tacotron2 [4], while the residual encoder has the same structure as baseline model except that we use bilingual PPG as input. All the weights in Equation (2) are empirically set to 1.0. For the transformation model, two FC layers with dropout 0.5 containing 512 hidden units and four BLSTM layers with 256 units per direction are used. This outputs are mapped to mel spectrograms by another FC layer. The text-to-bilingual PPG model and transformation model are separately trained.

5.2. Evaluation and analysis

To evaluate the code-switched speech synthesis performance of the baseline and proposed systems, we use each system to synthesize

[1] "https://github.com/fatchord/WaveRNN"

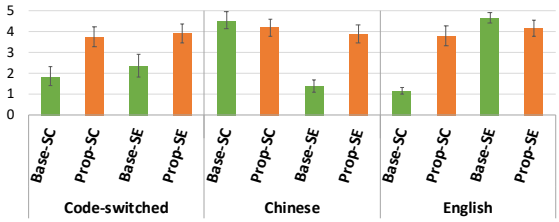

Fig. 3. MOS results on speech intelligibility for baseline (Base) and proposed (Prop) systems with code-switched, Chinese, and English input in Mandarin speaker's S_C or English speaker's S_E voice.

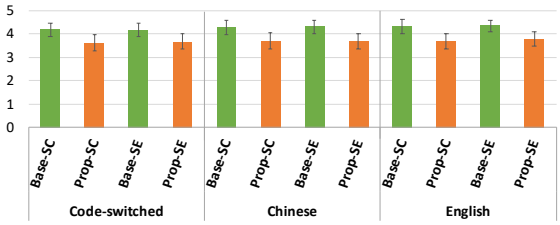

Fig. 4. MOS results on audio fidelity for baseline (Base) and proposed (Prop) systems with code-switched, Chinese, and English input in Mandarin speaker's S_C or English speaker's S_E voice.

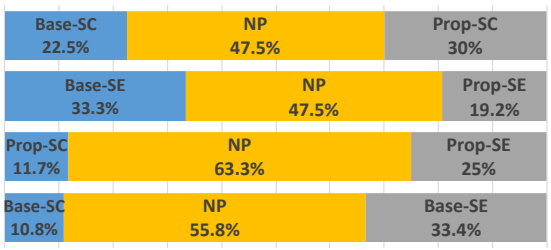

Fig. 5. Preference test results on speaker consistency for baseline (Base) and proposed (Prop) systems with code-switched input in Mandarin speaker's S_C or English speaker's S_E voice. NP denotes no preference.

audio samples with English, Chinese and code-switched text in both English speaker S_E's and Mandarin speaker S_C's voices for perceptual listening tests. Two mean opinion score (MOS) tests and one AB preference test are conducted for speech intelligibility, audio fidelity and speaker consistency respectively. 20 utterances for each setting are randomly chosen from the test set[2]. 17 native Mandarin speakers who are proficient in English participate in the listening tests.

Speech intelligibility. In the MOS test, the subjects listen to each pair of four utterances synthesized by the two systems with both speakers' voices. They are asked to give a 5-point scale score (5:excellent, 4:good, 3:fair, 2:poor, 1:bad) of speech intelligibility. The MOS result of speech intelligibility, presented in Fig. 3, shows that our proposed model with bilingual PPG is capable of synthesizing speech with stable intelligibility despite of changes in text input and speaker identity. This validates that bilingual PPGs benefit cross-lingual and code-switched speech synthesis as an intermediate feature, capturing phonetic information of both languages in a speaker-normalized space. Although the baseline model can generate very intelligible speech with Chinese input in S_C's voice and English input in S_E's voice, the speech intelligibility degrades seriously in cross-lingual and code-switched settings. Listeners comment that some samples sound like babbling. This indicates that the speaker characteristics and spoken language characteristics disentanglement are not well tackled in the baseline system. The language embedding does not capture sufficient language-dependent information for cross-lingual speech synthesis.

[2]Some samples are available in "https://csttsdemo.github.io/bppgCSTTS/"

Audio fidelity. Another MOS test is conducted similar to the one described above, except that listeners are asked to evaluate the audio fidelity of presented speech samples. The MOS result of audio fidelity is shown in Fig. 4. Although both systems can achieve decent audio fidelity, the baseline outperforms the proposed system in terms of audio fidelity in all settings. A possible reason is that the bilingual PPGs extracted in our experiments are not accurate enough. The numbers of senones in the English and Mandarin SI-ASR models are constrained to be small, degrading recognition performance (21.8% phone error rate for English ASR and 16.5% character error rate for Mandarin ASR). Our preliminary experiments show that when using PPGs with higher dimensions, the generated speech is even worse. There is a trade-off between the ASR performance and PPG dimensions, which will be further investigated in our future work.

Speaker consistency. Since we are most concerned with voice consistency within code-switched utterances, an AB preference test is conducted. Speech samples in AB preference test are generated by the baseline and proposed systems with code-switched input in both speakers' voices. Paired speech samples (A and B) with the same textual content from different settings with different speaker identities are presented to listeners. The listeners are required to provide a speaker consistency choice among 3 options: 1) sample A has greater speaker consistency within an utterance; 2) no preference (NP); 3) sample B has greater speaker consistency. Fig. 5 shows that there is no significant difference between the baseline and the proposed models in speaker identity preservation of code-switched speech synthesis. Neither the baseline nor the proposed model has much difference between preserving S_E's and S_C's voices, which reflects the effectiveness of speaker embedding in both systems.

6. CONCLUSION

In this paper, we propose a bilingual PPG based approach for code-switched TTS using only monolingual corpora. Bilingual PPGs, obtained by concatenating monolingual PPGs from English SI-ASR and Mandarin SI-ASR, are regarded as a bridge across speakers and language boundaries. Therefore, an attention-based encoder-decoder model is trained to map input text to bilingual PPGs, and the bilingual PPGs are mapped to mel spectrograms frame-wise by a BLSTM based model conditioned on speaker embedding. Experiments confirm the effectiveness of the proposed approach in synthesizing code-switched speech with decent speech intelligibility, audio fidelity and speaker consistency. Further improving the audio fidelity will be our future work.

7. ACKNOWLEDGEMENTS

This work is supported by the Joint Research Scheme of National Natural Science Foundation of China - Research Grants Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15).

References

[1] C. Myers-Scotton, *Duelling languages: Grammatical structure in codeswitching*, Oxford University Press, 1997.

[2] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] S. Arık, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. Neural Information Processing Systems (NIPS)*, 2017.

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

[5] Y. Wang, R. J. 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," in *Proc. Interspeech*, 2017.

[6] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[7] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft mulan-a bilingual tts system," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003.

[8] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (mandarin–english) tts," *IEEE Transactions on Audio, Speech, and Language Processing*, 2009.

[9] S. Sitaram, S. K. Rallabandi, and S. Black, "Experiments with cross-lingual systems for synthesis of code-mixed text," in *9th ISCA Speech Synthesis Workshop*, 2015.

[10] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009.

[11] K. Kintzley, A. Jansen, and H. Hermansky, "Event selection from phone posteriorgrams using matched filters," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[12] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *International Conference on Multimedia and Expo (ICME)*, 2016.

[13] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance.," in *Proc. Interspeech*, 2018.

[14] L. Liu, Z. Ling, Y. Jiang, M. Zhou, and L. Dai, "Wavenet vocoder with limited training data for voice conversion.," in *Proc. Interspeech*, 2018.

[15] L. Badino, C. Barolo, and S. Quazza, "Language independent phoneme mapping for foreign tts," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[16] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[17] F. L. Xie, F. K. Soong, and H. Li, "A kl divergence and dnn approach to cross-lingual tts," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[18] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams.," in *Proc. Interspeech*, 2016.

[19] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[20] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2005.

[21] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "Voice conversion-based multilingual to polyglot speech synthesizer for indian languages," in *International conference of IEEE Region 10 (TENCON)*, 2013.

[22] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text," in *Proc. Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.

[23] J. He, Y. Qian, F. K. Soong, and S. Zhao, "Turning a monolingual speaker into multilingual for a mixed-language tts," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

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

[25] Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[26] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural tts system with only monolingual data," *arXiv preprint arXiv:1904.06063*, 2019.

[27] S. Rallabandi and A. Black, "Variational attention using articulatory priors for generating code mixed speech using monolingual corpora," *Proc. Interspeech*, 2019.

[28] 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," *arXiv preprint arXiv:1907.04448*, 2019.

[29] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015.

[30] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, et al., "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[31] S. King and V. Karaiskos, "Blizzard challenge 2011," *Proc. Blizzard Challenge workshop*, 2011.

[32] "Chinese Standard Mandarin Speech Corpus," https://www.data-baker.com/open_source.html.

[33] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[34] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, 2011.