INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving Bilingual TTS Using Language And Phonology Embedding With Embedding Strength Modulator

Fengyu Yang, Jian Luan, Meng Meng, Yujun Wang

Xiaomi AI Lab, Beijing, China

{yangfengyu1, luanjian, mengmeng, wangyujun}@xiaomi.com

Abstract

In most cases, bilingual TTS needs to handle three types of input scripts: first language only, second language only, and second language embedded in the first language. In the latter two situations, it is a big challenge to accurately model the pronunciation and intonation of the second language in different contexts without mutual interference. This paper builds a Mandarin-English TTS system to acquire more standard spoken English speech from a monolingual Chinese speaker. We introduce phonology embedding to capture the English differences between different phonology with embedding masks. An embedding strength modulator is specially designed to capture the dynamic strength of language and phonology. Experiments show that our approach can produce significantly more natural and standard spoken English speech of the monolingual Chinese speaker. From analysis, we find that suitable phonology control contributes to better performance in different scenarios.
Index Terms: bilingual, speech synthesis, phonology, embedding mask, embedding strength modulator

1. Introduction

Nowadays, a bilingual text-to-speech(TTS)[1] system is necessary for many application scenarios like voice assistant. For example, the names of English songs and movies are often directly embedded in Chinese responses. A straightforward way to build a bilingual TTS system is by collecting speech data from bilingual speakers. [2] proposed a shared hidden Markov model (HMM)-based bilingual TTS system, using a Mandarin-English corpus recorded by a bilingual speaker. [3] presented a TTS system using a speaker and language factorized deep neural network(DNN) with a corpus of three bilingual speakers. However, mixed-lingual corpora are scarce while a large number of monolingual corpora are easily accessible.

Another way is to leverage monolingual speech data from different speakers [4, 5, 6, 7, 8, 9, 10]. [7] proposes a polyglot synthesis method adapting the shared HMM states to the target speaker, trained on monolingual corpora. [8] proposes to factorize speaker and language based on an HMM-based parametric TTS system. [9] utilizes a combined phonetic space in two languages to build a code-switched TTS system based on HMM. [10] maps the senones between two monolingual corpora in two languages with a speaker-independent DNN ASR output based on HMM TTS.

End-to-end TTS systems also extend to multilingual tasks using monolingual speech[11, 12, 13, 14, 15, 16, 17, 18]. [15] used Unicode bytes as a unified new language representation for multilingual TTS. 125 hours of speech were used and their system can read code-switching text, despite the problem of speaker inconsistency when cross-language. [16] trained with designed loss terms preserving the speaker's identity in multiple languages based on the VoiceLoop architecture [19]. The trained speech is recorded by 410 monolingual speakers speech from English, Spanish and German. [17] used an adversarial loss term to disentangle speaker identity from the speech content, which trained with 550 hours of speech from 92 monolingual speakers. Limited by corpus size, [18] proposed tone embedding and tone classifier for tone preservation to generate utterances in a proper prosodic accent of the target language.

Generally, each speaker speaks only one language, leading to speaker and language characteristics being highly correlated. Using only monolingual corpora for bilingual or multilingual TTS easily leads to heavy accent carry-over in synthesized speech or inconsistent voice between languages. Actually, bilingual corpus helps deal with the problem. [20] trained a TTS system transforming speaker embedding between languages from a bilingual speaker to other monolingual speakers for a high degree of naturalness. In this paper, we expect to utilize scarce bilingual corpora to acquire more standard spoken English from a monolingual speaker, which is highly correlated with phonology learning.

For example, in mixed-lingual utterances, the pronunciation of English by a non-native speaker, like Chinese, is strongly influenced by their native language and is most often different from the standard English pronunciation[21]. Mandarin derives pronunciation directly from the spellings of the word with different tones, which have a high grapheme-to-phoneme(g2p) correlation. In contrast, English is an alphabetic and highly non-phonemic language. In Consequence, native phonemic language speakers, whose pronunciation is influenced by the spelling of the word, often pronounce English words differently from standard English speakers[22]. In mixed-lingual utterances, these speakers, despite qualified bilingual speakers, generally replace some English phonemes with the closest phoneme in their native language, resulting in mispronunciation and differences in phonology like articulation change and intonation variation[23].

Given these challenges, building a state-of-the-art bilingual TTS system requires special designs handling the English differences in phonology between mixed-lingual and monolingual utterances. In this paper, our contributions include: (1) introducing phonology embedding to capture the English differences between mixed-lingual and monolingual utterances; (2) proposing embedding mask to language embedding for distinguishing information between different languages and phonology embedding for focusing on expression between different phonology of English; (3) designing embedding strength modulator(ESM) to capture the dynamic information of language and phonology, which helps to generate more standard spoken English speech; (4) experiments showing that static and dynamic

10.21437/Interspeech.2023-709

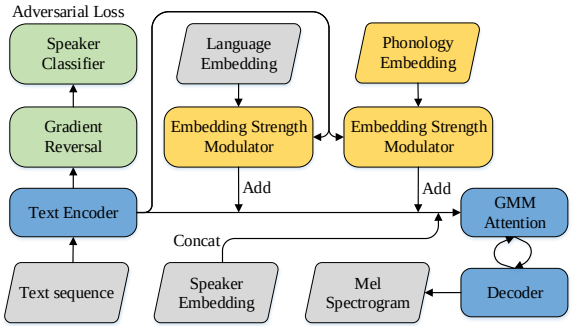


Figure 1: *Overview of the proposed bilingual architecture with specially designed modules marked in yellow color.*

components in ESM can control different attributes of phonology. Phonology decomposition and control can make a contribution to more standard spoken English expression and better performance in different scenarios.

2. Model Structure

Fig. 1 illustrates the proposed bilingual TTS architecture. The encoder-attention-decoder backbone with speaker and language embedding will be described in Sec. 2.1. The phonology embedding and specially designed masks for language and phonology embedding respectively will be described in Sec. 2.2. The embedding strength modulator will be described in Sec. 2.3.

2.1. Baseline

Our baseline system adopted from [24] is a popular Tacotron2[24]-based multilingual TTS architecture. It uses attention to bridge encoder and decoder. Language and speaker information are embedded in separate look-up tables. They are combined with the encoder output to distinguish different languages and speakers. Besides, an adversarially-trained speaker classifier is employed to disentangle text encoder output from speaker information. Mel-Lpcnet adopted from [25] is used as a vocoder to reconstruct waveform from given mel-spectrogram.

The architecture takes phoneme sequences as inputs for both English and Mandarin. Their phoneme sets are simply concatenated and no phoneme is shared across. Tone or stress tokens are inserted into the phoneme sequence at the end of each syllable. For Mandarin, there are 4 lexicon tones and one neutral tone. Instead, there are 4 stress types for English includes the sentence, primary, secondary, and none. Moreover, prosodic break tokens are inserted into the input sequence as well. Finally, the expanded phoneme set contains: 73 Mandarin phonemes, 39 English phonemes, 5 Mandarin tones, 4 English stresses, Mandarin character boundary, English syllable boundary, English liaison symbol and 4 shared prosodic break types, i.e. prosodic word (PW), prosodic phrase (PPH), intonation phrase (IPH) and silence at the beginning or end.

2.2. Embedding mask

Fig. 2 shows an example of embedding mask in language and phonology embedding.

Instead of broadcasting language embedding to all the tokens of the input sequence, the proposed method applies language embedding only to the token types shared across languages, i.e. PW, PPH, IPH and /sil/. Because other token types are language-specific already and need no additional information to distinguish language.

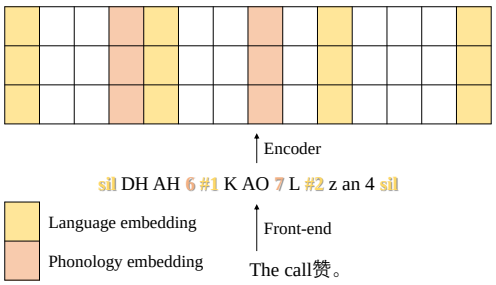


Figure 2: *An illustration of how to mask embedding. Language and phonology embedding only applied to the highlighted position of encoder outputs. The symbols #1, #2, #3 and /sil/ denote 4 shared prosodic break types. The numbers 1-5 denote tones of the previous Chinese syllable. The numbers 6-9 denote stresses of the previous English syllable.*

On the other hand, to capture the English differences between the mixed-lingual and monolingual utterances, a special phonology embedding is designed. To focus on English expression, it is applied to all English-specific tokens, including 4 types of stresses, syllable boundary and liaison symbol.

2.3. Embedding strength modulator

Even though the language and phonology embedding have been limited to only part of input tokens by masks, we think their strength should vary for different contexts. To capture the dynamic strength of languages and phonology, we propose an attention-based embedding strength modulator, whose framework is similar to [26, 27, 28].

The structure of the ESM is shown in Fig. 3. There are two sub-networks in ESM: multi-head attention and a feed-forward network. The layer normalization and residual connection are applied to both of the sub-networks. Formally, from the encoder output with scaled positional encoding E_o, and the language or phonology embedding LP, the first sub-network M_o and the second sub-network F_o are calculated as:

$$M_o = \mathrm{MH}(E_o,\ \mathrm{LN}(LP),\ \mathrm{LN}(LP)) + LP, \tag{1}$$

$$F_o = \mathrm{FFN}(\mathrm{LN}(M_o)) + M_o. \tag{2}$$

where MH(query, key, value), FFN(·) and LN(·) are multi-head attention, feed-forward network and layer normalization respectively. Since the attention key and value (LP) have only one item, the energy need not be normalized by softmax operation. Instead, each head in multi-head attention is computed by:

$$head_h = \alpha_h \cdot V_h = \frac{Q_h \cdot K_h}{\|Q_h\| \|K_h\|} \cdot V_h, \tag{3}$$

where $\|\cdot\|$ is the L2 norm of the last dimension, $\{Q, K, V\}$ represent query, key and value through linear transformation respectively and the strength α is a scaled cosine similarity between the query and key to be in the range of [-1, 1].

In particular, there are two components in Fig. 3 marked in yellow color. The original embedding learned for each language and phonology is regarded as a static component. While the output of multi-head attention, the static embedding multiplied by a dynamic weight, is regarded as a dynamic component. We will analyze the roles each component of language and phonology embedding play in Sec. 3.3.

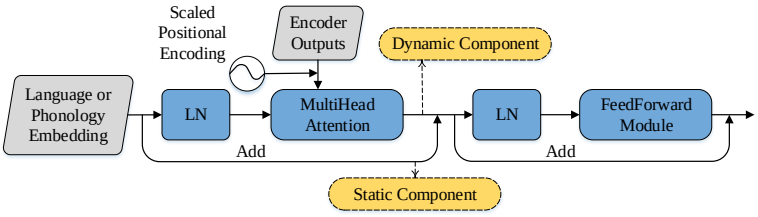


Figure 3: *The structure of the embedding strength modulator of both language and phonology embedding.*

3. Experiments

3.1. Basic setups

Models are trained with proprietary datasets composing three kinds of high-quality speech: (1) bilingual corpus from two Chinese speakers, 45000 and 25000 Mandarin utterances for female and male speakers respectively, 9000 mixed-lingual utterances and 9000 English utterances for both speakers; (2) English corpus from two American speakers, 9000 and 25000 English utterances for female and male speakers respectively; (3) Mandarin corpus from a female Chinese speaker, 9000 Mandarin utterances for cross-lingual experiments.

Labels of the above corpora in language and phonology embedding are described in Tab. 1. Mandarin utterances are all from Chinese speakers. Their language is labeled Mandarin and no English phonology label is required. For plain English utterances, the corpora recorded by both American and Chinese speakers are labeled as English for language and Standard-English for phonology. These Chinese speakers pronounce English well and the corpora recording by American speakers are used as supplementary English datasets and are beneficial to speaker learning. Particularly, since English parts in mixed-lingual utterances are in a small amount and are mostly words and abbreviations with heavy Chinese phonology, we label them Mandarin for language and Chinese-English for phonology, treated as Mandarin utterances.

The additional inputs of the learned speaker (64-dim), language and phonology embedding (both 512-dim same with the dimensions of encoder output) are injected into the backbone. In ESM, the first sub-network includes 8-head multi-head attention and the feed-forward sub-network consists of two convolution networks with 2048 and 512 hidden units. Linguistic inputs have been introduced in Sec. 2.2 and for acoustic features, we use an 80-band mel-spectrogram extracted from 16kHz waveforms. We built the following systems for comparison:

- **BASE**: Baseline system with sentential language embedding as described in Sec. 2.1;
- **EM**: Baseline system with specially designed language and phonology embedding as described in Sec. 2.2;
- **ESM**: Baseline system with specially designed language and phonology embedding through ESM as described in Sec. 2.3.

3.2. Subjective evaluation

We conduct Mean Opinion Score (MOS) evaluations of speech naturalness and speaker similarity via subjective listening tests. 20 speakers are asked to listen to the generated 20 English utterances and 10 mixed-lingual utterances. MOS results are reported in Tab. 2. Except for parts of samples in listing tests, generated Mandarin demos of this monolingual speaker are also shown in demo pages[1].

[1]Samples can be found from: https://fyyang1996.github.io/esm/

We can find that the EM system with masked embedding brings better performance on both speech naturalness and speaker similarity than the conventional BASE system. It indicates that masked embedding captures features that better represent language and phonology. For the further proposed embedding strength modulator, we find that by capturing the dynamic strength of language and phonology system ESM achieves significantly better performance than the EM system. It demonstrates that the dynamic strength of language and phonology is beneficial to speech naturalness and speaker similarity of generated speech.

3.3. ESM component analysis

As mentioned above, the output of ESM may be regarded as the combination of a static component and a dynamic component. One simple method of analyzing the contribution of each component is to condition the model on only one component at each time. In the generation phase, we replace the static or dynamic component from Mandarin label to English label for language embedding or from Chinese-English phonology label to Standard-English phonology label for phonology embedding respectively. Fig. 4 shows the spectrogram and F0 contour, extracted by parselmouth[29], of the same sentence synthesized with six kinds of label combinations as described below:

(a) Base combination: using Mandarin and Chinese-English phonology labels both in dynamic and static components;

(b) Reference combination: using English and Standard-English phonology labels both in dynamic and static components;

(c) Based on (a), replacing dynamic phonology embedding from Chinese-English to Standard-English phonology.

(d) Based on (a), replacing static phonology embedding from Chinese-English to Standard-English phonology;

(e) Based on (a), replacing dynamic language embedding from Mandarin to English;

(f) Based on (a), replacing static language embedding from Mandarin to English;

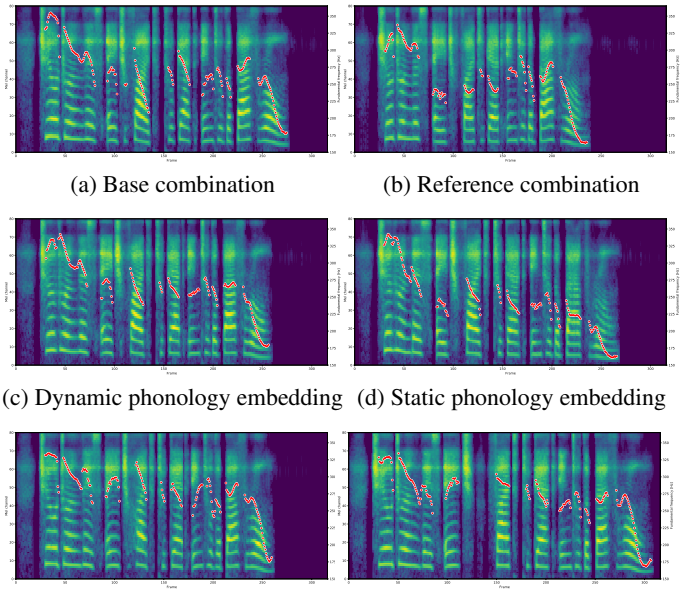
Empirically, we find that each component represents articulation, intonation, speaking rate and pause duration changes respectively, which influence phonology collectively. Listing to the samples of (a) and (c) in the demo page, we can easily hear about articulation changes between them, which is difficult to be caught sight of. Perceptually, the trend of F0 values in Fig. 4(d) is different from that in Fig. 4(a), showing that static phonology embedding major affects intonation. Fig. 4(e) shows that replacing the dynamic language embedding from Mandarin to English causes a gradual compression of the spectrogram and F0 values in the time domain. We believe that the dynamic language embedding encodes the information correlated with speaking rate variation. Besides, syllables in Fig. 4(f) have distinct intervals compared with that in Fig. 4(a), which demonstrates that static

Table 1: *Labels of trained corpora in language and phonology embedding.*

Corpus			Language embedding	Phonology embedding
(1) Chinese speaker	Train	Mandarin	Mandarin	None
		Mixed-lingual	Mandarin	Chinese-English
		English	English	Standard-English
(2) American speaker	Train	English	English	Standard-English
(3) Chinese speaker	Test	Mandarin	Mandarin	None

Table 2: *The MOS of different systems with confidence intervals of 95%.*

Model	BASE	EM	ESM
Naturalness	3.81±0.12	4.03±0.10	**4.39±0.08**
Similarity	3.79±0.12	3.91±0.11	**4.04±0.10**



(a) Base combination (b) Reference combination

(c) Dynamic phonology embedding (d) Static phonology embedding

(e) Dynamic language embedding (f) Static language embedding

Figure 4: *Spectrogram and F0 of a test sentence generated by different combinations, which refers to 2.1 in demo page.*

language embedding represents the average duration of pauses. More demos and be found in the demo page.

3.4. Control

To validate the above analysis, we conduct MOS evaluations of speech naturalness and speaker similarity via subjective listening tests. 20 speakers are asked to listen to the generated 15 English utterances for enhancing English expressiveness and 15 mixed-lingual utterances for smooth mixed-lingual transition. Demos can be found in 3 and 4 on the demo pages.

Enhance expressiveness To enhance the expressiveness of a plain English text, we double the dynamic components of both language and phonology embedding while remaining their static components. The "double" herein means that the final vector has a double distance of the reference vector to the base vector. For language embedding, the reference is English and the base is Mandarin. While for phonology embedding, the reference is Standard English and the base is Chinese-English. Fig. 5 shows the results of MOS evaluations. We find that by the

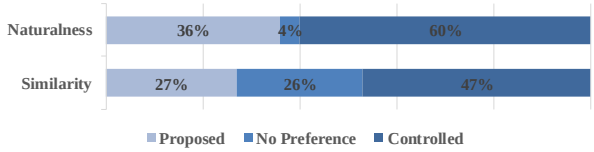


Figure 5: *A/B preference results for control in enhancing expressiveness or not with confidence intervals of 95% and p-value<0.0001 from the t-test.*

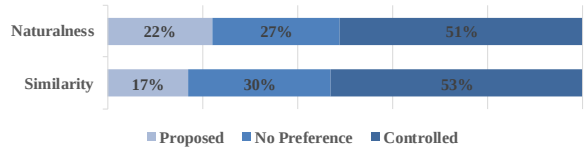


Figure 6: *A/B preference results for control in smooth transition or not with confidence intervals of 95% and p-value<0.0001 from the t-test.*

"double" operation herein system ESM achieves significantly better performance than the ESM system on speech naturalness. It indicates that the control operation enhances English expressiveness significantly.

Smooth transition When synthesizing a mixed-lingual text, we modify the language labels of embedded English words from Mandarin to English while phonology labels of them from Chinese-English to Standard-English. Particularly, their static component of phonology embedding remains Chinese-English. In this way, the English words will have standard-English articulation but more compatible intonation with the context of Chinese words. Fig. 6 shows the results of MOS evaluations. It can be found that the controlled ESM system brings better performance on both speech naturalness and speaker similarity than the proposed ESM system. It demonstrates that the control operation is beneficial to smooth mixed-lingual transition.

4. Conclusions

This paper builds a Mandarin-English TTS system for a monolingual Chinese speaker. We introduce phonology embedding and a special designed mask for language and phonology embedding. They are employed to distinguish two languages and the English phonological differences between monolingual and embedded cases respectively. Furthermore, the proposed embedding strength modulator enables language and phonology embedding to be variable with token context. Experiments show that our approach can produce significantly more natural and standard spoken English speech than baseline. Ablation analysis on different components demonstrates that English phonology can be tuned effectively for various scenarios.

5. References

[1] M. Chu, H. Peng, Y. Zhao, Z. Niu, and E. Chang, "Microsoft mulan-a bilingual tts system," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2003, pp. 264–267.

[2] 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 (TASLP)*, vol. 17, no. 6, pp. 1231–1239, 2009.

[3] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in dnn-based tts synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5540–5544.

[4] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an hmm-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[5] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis," in *2016 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2016, pp. 2468–2472.

[6] I. Himawan, S. Aryal, I. Ouyang, S. Kang, P. Lanchantin, and S. King, "Speaker adaptation of a multilingual acoustic model for cross-language synthesis," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7629–7633.

[7] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2005, pp. I–1.

[8] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 20, no. 6, pp. 1713–1724, 2012.

[9] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. W. Black, "Experiments with cross-lingual systems for synthesis of code-mixed text." in *SSW*, 2016, pp. 76–81.

[10] F.-L. Xie, F. K. Soong, and H. Li, "A kl divergence and dnn approach to cross-lingual tts," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5515–5519.

[11] 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 *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6935–6939.

[12] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural tts system with only monolingual data," in *2019 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2019.

[13] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7619–7623.

[14] S. Zhao, T. H. Nguyen, H. Wang, and B. Ma, "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," in *2020 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2020, pp. 2927–2931.

[15] 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 *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5621–5625.

[16] E. Nachmani and L. Wolf, "Unsupervised polyglot text-to-speech," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7055–7059.

[17] 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 *2019 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2019, pp. 2080–2084.

[18] R. Liu, X. Wen, C. Lu, and X. Chen, "Tone learning in low-resource bilingual tts." in *2020 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2020, pp. 2952–2956.

[19] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *2018 International Conference on Learning Representations (ICLR)*, 2018.

[20] S. Maiti, E. Marchi, and A. Conkie, "Generating multilingual voices using speaker space translation based on bilingual speaker data," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7624–7628.

[21] Y. Lee, S. Shon, and T. Kim, "Learning pronunciation from a foreign language in speech synthesis networks," *arXiv preprint arXiv:1811.09364*, 2018.

[22] A. Baby, P. Jawale, S. Vinnaitherthan, S. Badam, N. Adiga, and S. Adavanne, "Non-native english lexicon creation for bilingual speech synthesis," *arXiv preprint arXiv:2106.10870*, 2021.

[23] J. He, Y. Qian, F. K. Soong, and S. Zhao, "Turning a monolingual speaker into multilingual for a mixed-language tts," in *2012 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2012, pp. 963–966.

[24] 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 *2018 IEEE international conference on acoustics, speech and signal processing 1(ICASSP)*. IEEE, 2018, pp. 4779–4783.

[25] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[26] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *2020 International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 10 524–10 533.

[27] F. Yang, J. Luan, and Y. Wang, "Improving emotional speech synthesis by using sus-constrained vae and text encoder aggregation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8302–8306.

[28] F. Yang, S. Yang, Q. Wu, Y. Wang, and L. Xie, "Exploiting deep sentential context for expressive end-to-end speech synthesis," *arXiv preprint arXiv:2008.00613*, 2020.

[29] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.