END-TO-END CODE-SWITCHED TTS WITH MIX OF MONOLINGUAL RECORDINGS

Yuewen Cao[1]*, Xixin Wu*[1]*, Songxiang Liu*[1]*,Jianwei Yu*[1]*, Xu Li*[1]
Zhiyong Wu[*1,2]*, Xunying Liu*[1]*, Helen Meng*[1]

[1] Human-Computer Communication 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
{ywcao, wuxx, sxliu, jwyu, xuli, zywu, xyliu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

State-of-the-art text-to-speech (TTS) synthesis models can produce monolingual speech with high intelligibility and naturalness. However, when the models are applied to synthesize code-switched (CS) speech, the performance declines seriously. Conventionally, developing a CS TTS system requires multilingual data to incorporate language-specific and cross-lingual knowledge. Recently, end-to-end (E2E) architecture has achieved satisfactory results in monolingual TTS. The architecture enables the training from one end of alphabetic text input to the other end of acoustic feature output. In this paper, we explore the use of E2E framework for CS TTS, using a combination of Mandarin and English monolingual speech corpus uttered by two female speakers. To handle alphabetic input from different languages, we explore two kinds of encoders: (1) shared multilingual encoder with explicit language embedding (LDE); (2) separated monolingual encoder (SPE) for each language. The two systems use identical decoder architecture, where a discriminative code is incorporated to enable the model to generate speech in one speaker's voice consistently. Experiments confirm the effectiveness of the proposed modifications on the E2E TTS framework in terms of quality and speaker similarity of the generated speech. Moreover, our proposed systems can generate controllable foreign-accented speech at character-level using only mixture of monolingual training data.

Index Terms— code-switching, multilingual speech synthesis, end-to-end, foreign accent

1. INTRODUCTION

Due to the rise of globalization, alternating of languages in text or speech, referred to as code-switching (CS) in linguistics [1], is a common phenomenon in social media text, informal messages and voice navigation. These situations happen to be the scenarios where text-to-speech (TTS) systems are widely used as speech interface. It is desirable to synthesize such code-switched sentences with a consistent and natural voice. However, current TTS systems mostly assume that the inputs are in a single language. The consequence is that they often synthesize incorrect pronunciations or even miss the words when the language changes in the sentence.

Ideally, we need CS speech data for all languages from a *single* multilingual speaker to train a code-switched speech synthesizer. However, it is not easy to find such a multilingual speaker to record CS data in large quantities. Most current work utilizes a combination of widely available monolingual data. Existing code-switched TTS systems are mostly concatenative or HMM-based, and can be categorized into three main types: (i) Polyglot systems that create a combined phone set covering all languages. For example, [2] proposed a combined diphone inventory using data from a multilingual speaker. [3] combined monolingual data from different speakers to train an average voice, and then adapted the average voice to the target language and speaker. (ii) Multilingual TTS systems that use separated front-ends to process the text portions of different languages while sharing a unit-selection module [4] or acoustic model [5, 6]. (iii) Unit mapping methods bridge the language gap by substitution, e.g., through frame mapping [7], state mapping [8] and senone mapping [9], etc. Statistical information extracted from the training data was adopted for modifying the pitch and duration of CS speech [7, 10, 11].

Despite the success of previous code-switched TTS systems, they rely on hand-made front-ends, which are developed by language-specific experts. This greatly increases the effort required to develop TTS systems, especially for a new language. Besides, specific criterion is needed to construct the cross-language mapping relationship, e.g., Kullback-Leibler divergence [8, 9], trajectory tiling approach [7], etc. Strong accented voice will be caused for lack of appropriate substitutions of phonetic units between languages. Disjointly optimizing each stage in the system may lead to accumulation of errors.

With easy access to large-scale monolingual corpora nowadays, the end-to-end (E2E) architecture has proven its effectiveness in monolingual TTS synthesis [12, 13], generating natural speech with high quality. The E2E TTS system directly converts character input sequences to audio samples or acoustic outputs, therefore an implicit grapheme-to-phoneme model is learned. This design alleviates the need of domain expertise and jointly training the whole model can hopefully reduce the compounded errors. Such a system is also flexible to incorporate various conditioning attributes to control the model output, e.g., speaker codes [14], style tokens [15, 16], etc. Besides, the E2E architecture has also been successfully applied to code-switched tasks suck as speech recognition [17].

In this paper, we explore the efficacy of using the Tacotron-based E2E framework [18], which is composed of an encoder and a decoder with attention, for code-switched TTS using a combination of a Mandarin and an English monolingual speech corpus uttered by

*Corresponding author

two female speakers. To handle code-switched alphabetic text input, we explore two kinds of encoders: (1) shared multilingual encoder with explicit language embedding (LDE) for alphabetic sequences in different languages; (2) separated monolingual encoder (SPE) to encode each language. Since we want the model to generate speech in one speaker's voice consistently, a discriminative code is incorporated into the decoder and is used for both systems to control the output speech timbres. Experiments confirm the effectiveness of the two proposed systems in terms of quality and speaker similarity of the generated speech. Moreover, our proposed systems have the flexibility to adjust the degree of perceived accent at character level, even though only non-accented training data is used. To the best of our knowledge, the current paper is among the first to use the E2E architecture for code-switched speech synthesis using the mix of monolingual data. Our systems have several advantages:

- They relax the data constraint on bilingual speaker's recordings for code-switched speech synthesis.
- They learn the language-specific grapheme-to-phoneme knowledge automatically through jointly training the whole model, alleviating the need for language domain-expert knowledge and hand-crafted engineering.
- They have the flexibility to generate controllable degree of foreign-accented speech without using accented training data.

The rest of the paper is organized as follows: Section 2 introduces the two code-switched TTS systems. Section 3 describes the experiments details. Section 4 presents the evaluation results. The conclusion is drawn in section 5.

2. END-TO-END CODE-SWITCHED TTS SYSTEMS

2.1. Code-switched TTS based on language embedding

The input text is first processed by the text-normalization module, which changes numbers (date, money, etc.) and abbreviations into readable text strings. Then we identify the language of each word in the sentence using a simple orthography-based method. The Chinese characters in normalized text are then converted to Pinyin with tonal information. In this way, each sentence $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$ has a character-level languageID sequence $\boldsymbol{l} = \{l_1, l_2, ..., l_n\}$ where each element denotes the language ID of corresponding character, and a discriminative code $\boldsymbol{s}$ that captures speaker-related acoustic differences.

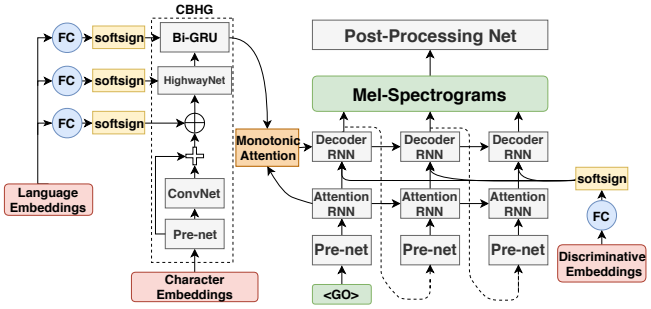

Fig. 1. The network architecture of the LDE system.

Since both Pinyin and English use alphabetic characters, the model needs to distinguish the phonetic sound of the same character from different languages. Although Mandarin is tonal and English is not, the model can learn partial linguistic knowledge from the Pinyin's encoded tone information. However, the default Tacotron is likely to synthesize speech with inconsistent voices with respect to different languages in CS text or even fails to generate intelligible speech. In order to better model the differences between languages and explicitly model language alternation at local context, we augment the encoder with explicit character-level languageID. As shown in Fig. 1, we incorporate the language information into multiple portions of encoder CBHG architecture. First, we obtain the language embedding $\boldsymbol{L}$ from the embedding lookup table by the languageID $\boldsymbol{l}$. The obtained language embedding is then mapped to high-level representations through three distinct fully-connected (FC) layers activated by softsign, before being incorporated into the shared multilingual encoder. To be specific, one of them is concatenated with the ConvNet outputs, one is used as the gate in each highway layer, and the remaining one is set as the initial state of the BiGRU. During training, the elements of the languageID $\boldsymbol{l}$ for each sentence $\boldsymbol{x}$ are actually all zeros or all ones. During testing, the languageID $\boldsymbol{l}$ is determined by the language boundary information in the CS text.

We use monotonic attention, which achieves soft-monotonic attention by training in expectation [19]. This approach forces attention to follow a monotonic pattern and enables the model to learn better alignments when the text input is code-switched. The remaining parts are the same as the default Tacotron, except that a discriminative embedding $\boldsymbol{S}$ is incorporated into the decoder RNN to control the output speech timbres. The discriminative embedding is obtained by performing discriminative code lookup, and is concatenated with previous time-step decoder output and context information before being sent to decoder RNN. This design enables the generated speech in a single speaker's voice. The language embedding and discriminative embedding are jointly learned with the model by back-propagation.

2.2. Code-switched TTS based on separate encoders

To further alleviate the mutual interference of inputs from different languages, we investigate the use of separate encoders for each language. As shown in Fig. 2, we have an English encoder $Encoder_{en}$ and a Mandarin encoder $Encoder_{ch}$. The input character sequence

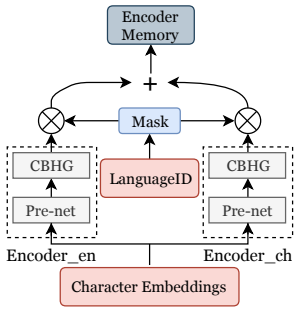

Fig. 2. The network architecture of the encoder in the SPE system.

$\boldsymbol{x}$ is sent to both encoders. Then, the languageID sequence $\boldsymbol{l}$ is used as a mask to fetch the language portions from respective encoder memory. The final encoder memory is:

$$Mem = Mem_{ch} \odot Mask_{ch} + Mem_{en} \odot Mask_{en}, \quad (1)$$

where each encoder memory (Mem_{en}, Mem_{ch}) is multiplied element-wisely with respective mask generated from languageID

sequence. During training stage, the language-specific encoder is actually only trained with corresponding monolingual data, where either $Mask_{en}$ or $Mask_{ch}$ is set to $\mathbf{1}$. In testing stage, the code-switched text obtain the corresponding encoder memory for each language portions as defined in Equation (1). We observe from our preliminary experiments that allowing each encoder to "see" the whole sentence gets better generation performance. Even though each encoder may mishandle text from the other language, maintaining the intact contextual information of language portions can help reduce the mismatch in language boundaries. This helps to improve alignments when synthesizing speech from CS text. The remaining part of the decoder in SPE is identical with that in LDE.

3. EXPERIMENTS

3.1. Experimental setup

In our experiment, one American English speech corpus [20] and one Mandarin speech corpus [21] uttered by two different female speakers are used. Both corpora are in neutral broadcast news reading style. We randomly select 5000 utterances (8 hours for Mandarin and 7 hours for English) from each corpus as training data. 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, as shown in Fig. 3. Alphabetic

(i) That's why 很多人都用地铁.
Translation: That's why many people use the subway.
(ii) 岳阳 Tower is one of the Three Great Towers of 江南.
Translation: Yueyang Tower is one of the Three Great Towers of Jiangnan.

Fig. 3. Examples of code-switched test data.

sequences of Pinyin and English are used as the input. The log-magnitude linear-scale spectrograms and 80-band Mel-scale spectrograms extracted with 50ms Hanning window, 12.5ms frame shift and 2048-point Fourier transform are used as the output.

3.2. Systems implementation

We use Tacotron (Tac) as baseline system and focus on investigating the LDE and SPE systems from two aspects: (i) whether the models are capable of generating natural speech with the CS alphabetic text input and (ii) whether the models can synthesize CS speech with a single voice consistently.

The Tacotron system is implemented by following [18]. For the LDE system, the language embedding is set to have dimension 32 and is mapped to higher dimension with a 64-unit FC layer activated by softsign. For the SPE system, the character-level language code is used as the mask to fetch respective language parts from two encoder memories. The 32-dimension discriminative embedding used for both the LDE and SPE systems, is projected the same way as language embedding. We clip gradients when their global norm exceeds 1 and use parallel-mode monotonic attention with initial energy function scalar bias set to -1. The Griffin-Lim algorithm [22] is used to convert the output spectrograms to waveforms for all systems. Three systems are compared: Tac, LDE and SPE. We synthesize audio samples with English, Chinese and code-switched text in two speakers' voice using each system.

4. EVALUATIONS AND ANALYSIS

4.1. Subjective evaluation

We conduct a 5-scale mean opinion score (MOS) test and a ABX test for speech naturalness and speaker similarity, respectively. Twenty native Mandarin speakers who are proficient in English are invited to participate in each test. Twenty utterances for each setting are randomly chosen from the test set [1]. In the MOS test, speech samples generated by the Tac, LDE and SPE systems with English (EN), Mandarin (CH) or code-switched (CS) text as input in both speakers' voices are presented to the listeners. The MOS result, which is illustrated in Fig. 4, shows that our proposed LDE and SPE systems are capable of generating more natural speech in both speakers' voices than the baseline when the input is code-switched. The MOS re-

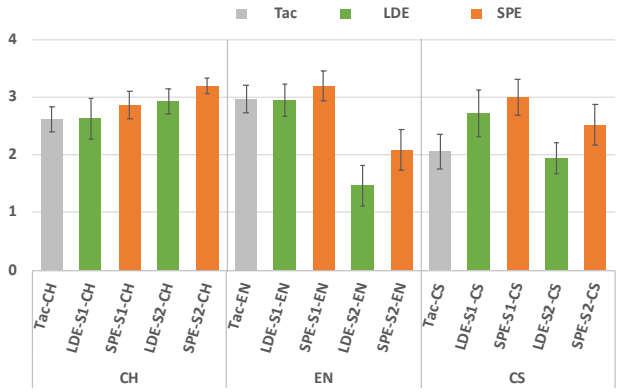

Fig. 4. MOS results of Tac, LDE, SPE systems with English (EN), Mandarin (CH) or code-switched (CS) input in English speaker (s1) or Mandarin speaker's (s2) voice.

sults of SPE-S1-CH, SPE-S1-EN and SPE-S1-CS show that we can synthesize natural Mandarin, English and code-switched speech in the American speaker's voice using the SPE system. Similar phenomenon can be seen for the LDE system. These indicate that both systems can handle alphabetic text input of different languages in the American speaker's voices. However, neither the SPE nor the LDE system performs well on English text using the Mandarin speaker's voice, indicated by MOS results of SPE-S2-EN and LDE-S2-EN. Possible reason is that we use Pinyin as input for Mandarin and alphabetic characters for English, where Pinyin contains richer phonetic information, such as tones, than English character sequence. Thus the decoder conditioned on discriminative code of the Mandarin speaker will fail to handle the encoder output lacking of tonal information. Besides, the SPE system outperforms the LDE system on all settings. It shows that the model benefits more from less language interference when encoding the inputs to hidden representations.

In the ABX test, speech samples are generated by the LDE and SPE systems with Mandarin input in the American speaker's voice, with English input in the Mandarin speaker's voice, as well as generated by the baseline with Mandarin and English input. The listeners are required to provide a speaker similarity choice among 3 options: 1) similar to the English speaker's voice; 2) no preference; 3) similar to the Mandarin speaker's voice. We use these setups to investigate

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

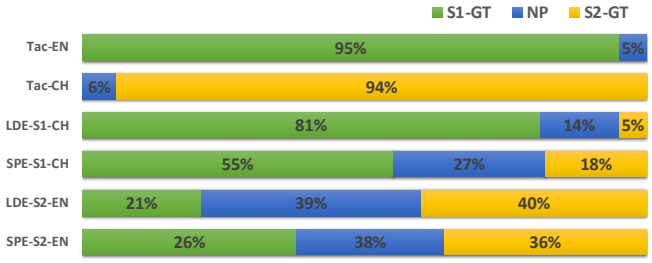

Fig. 5. Speaker similarity results of English (EN) and Mandarin (CH) speech generated by Tac system, Mandarin (CH) speech in English speaker's (S1) voice generated by LDE and SPE systems, English (EN) speech in Mandarin speaker's (S2) voice generated by LDE and SPE systems. S1-GT and S2-GT are the ground truth of English speaker and Mandarin speaker respectively. NP denotes no preference.

cross-language speaker similarity preservation of our systems. Fig. 5 shows that the baseline will synthesize different language in different voice. Both the LDE and SPE systems can outperform the baseline by successfully maintaining the American speaker's voice when synthesizing Mandarin speech. Neither systems can maintain the Mandarin speaker's voice well when synthesizing English speech, which may be partially due to the low quality of synthesized speech.

4.2. Foreign-Accent evaluation

We concentrate on American-English-accented Mandarin, since the code-switched systems conditioned on discriminative code of the American speaker can generate the three types of speech with the same voice and desirable quality. Also, we have easier access to native Mandarin subjects. For LDE, we first retrieve the English and Mandarin embedding vectors ($\boldsymbol{L_{en}}$, $\boldsymbol{L_{ch}}$) from the trained model, and manipulate the language embeddings by linear interpolation as:

$$\boldsymbol{L} = \boldsymbol{\alpha} \odot \boldsymbol{L}_{en} + (\mathbf{1} - \boldsymbol{\alpha}) \odot \boldsymbol{L}_{ch}, \tag{2}$$

where $\boldsymbol{L}$ is used for synthesis and $\boldsymbol{\alpha}$ is the accent coefficient. We expect a larger $\boldsymbol{\alpha}$ value will make a stronger American-English accent. For SPE, we use a similar linear interpolation between English and Mandarin encoder memories (Mem_{en}, Mem_{ch}) as:

$$Mem = \boldsymbol{\alpha} \odot Mem_{ch} \odot Mask_{ch} + (\mathbf{1} - \boldsymbol{\alpha}) \odot Mem_{en} \odot Mask_{en}, \tag{3}$$

where Mem is used for synthesis. The proposed SPE and LDE systems allow adjusting the degree of perceived accent in the generated speech at character level by the coefficient vector $\boldsymbol{\alpha}$, as shown in Equation (2) and (3).

We set $\boldsymbol{\alpha}$ to only contain 0, 0.5 or 1 to generate sentence-level accented audio samples in the American speakers' voice. A foreign accent degree judgement on a scale of 1 (native-like) to 7 (very strong accent) is conducted as in [23] to verify the control ability of the coefficient $\boldsymbol{\alpha}$. Twenty native Mandarin speakers are invited to evaluate 20 utterances for each setting. Fig. 6 shows that both the SPE and LDE systems can adjust the degree of perceived accent, while a larger $\boldsymbol{\alpha}$ makes a stronger accent. Also, SPE is more sensitive to the coefficient $\boldsymbol{\alpha}$ than LDE.

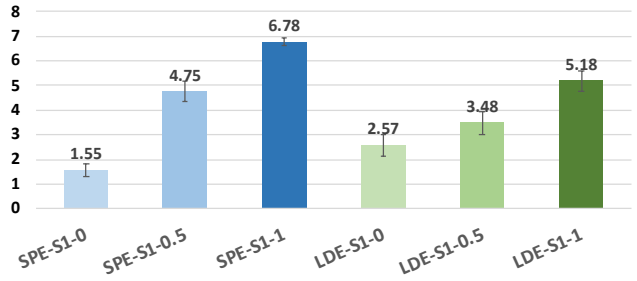

Fig. 6. Foreign accent ratings of Mandarin speech in English speaker's (s1) voice generated by SPE and LDE systems with accent coefficient $\mathbf{0, 0.5, 1}$. Larger value means stronger accent.

5. CONCLUSIONS

In this paper, we propose the LDE system and SPE system in E2E framework for code-switched TTS. Our systems have the following advantages:(1) They relax the requirement for multilingual data from the same speaker. (2) They can generate speech with the CS alphabetic text input in both speakers' voice. (3) They have the flexibility of adjusting the degree of perceived accent at the character level. Experiments verify the effectiveness of both systems, in synthesizing monolingual English or Mandarin, as well as code-switched speech. In the future, we will investigate techniques for transferring a trained model based on Mandarin speech to cover synthesizing English speech.

6. ACKNOWLEDGEMENTS

This work is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N CUHK404/15).

7. REFERENCES

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

[2] 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.

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

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

[5] K. R. Chandu, S. K. Rallabandi, S. Sitaram, and A. W. Black, "Speech synthesis for mixed-language navigation instructions," *Proc. Interspeech 2017*.

[6] H. Li, Y. Kang, and Z. Wang, "Emphasis: An emotional phoneme-based acoustic model for speech synthesis system," *arXiv preprint arXiv:1806.09276*, 2018.

[7] 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.

[8] H. Liang, Y. Qian, and F. K. Soong, "An hmm-based bilingual (mandarin-english) tts," *Proc. Speech Synthesis Workshop6 (SSW6)*, 2007.

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

[10] S. Rallabandi and A. W. Black, "On building mixed lingual speech synthesis systems," *Proc. Interspeech 2017*.

[11] Y. Zhang and J. Tao, "Prosody modification on mixed-language speech synthesis," in *Proc. Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2008.

[12] A. van den 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.

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[14] 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.

[15] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[16] X. Wu, Y. Cao, M. Wang, S. Liu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Rapid style adaptation using residual error embedding for expressive speech synthesis," *Proc. Interspeech 2018*.

[17] H. Seki, S. Watanabe, T. Hori, J. Le Roux, and J. R Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proc. Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[18] 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.

[19] C. Raffel, M. T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," *arXiv preprint arXiv:1704.00784*, 2017.

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

[21] L. Cai, D. Cui, and R. Cai, "Th-coss, a mandarin speech corpus for tts," *Journal of Chinese Information Processing*, 2007.

[22] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[23] M. L. G. Lecumberri, R. Barra-Chicote, R. P. Ramón, J. Yamagishi, and M. Cooke, "Generating segmental foreign accent," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.