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

Dynamic Soft Windowing and Language Dependent Style Token for Code-Switching End-to-End Speech Synthesis

Ruibo Fu[1,2], *Jianhua Tao*[1,2,3], *Zhengqi Wen*[1], *Jiangyan Yi*[1], *Chunyu Qiang*[1,2], *Tao Wang*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing

{ruibo.fu,jhtao,zqwen,jiangyan.yi,chunyu.qiang,tao.wang}@nlpr.ia.ac.cn

Abstract

Most of current end-to-end speech synthesis assumes the input text is in a single language situation. However, code-switching in speech occurs frequently in routine life, in which speakers switch between languages in the same utterance. And building a large mixed-language speech database is difficult and uneconomical. In this paper, both windowing technique and style token modeling are designed for the code-switching end-to-end speech synthesis. To improve the consistency of speaking style in bilingual situation, compared with the conventional windowing techniques that used fixed constraints, the dynamic attention reweighting soft windowing mechanism is proposed to ensure the smooth transition of code-switching. To compensate the shortage of mixed-language training data, the language dependent style token is designed for the cross-language multi-speaker acoustic modeling, where both the Mandarin and English monolingual data are the extended training data set. The attention gating is proposed to adjust style token dynamically based on the language and the attended context infromation. Experimental results show that proposed methods lead to an improvement on intelligibility, naturalness and similarity.

Index Terms: speech synthesis, code-switching, dynamic soft windowing, language dependent style token

1. Introduction

End-to-end speech synthesis, such as Tacotron, can achieve the state-of-art performance, and even close to human recording based on the monolingual corpus [1–5]. However, code-switching in speech occurs frequently in routine life, in which speakers switch between languages in the same utterance. The code-switching speech synthesis system is expected to generate high naturalness and similarity mixed-lingual utterance, which could be perceived as from one speaker [6]. There are several challenges to build a code-switching speech synthesis system. On the one hand, the speaking style could be inconsistent at the code-switching points between two different languages. On the other hand, building a large mixed-language speech database is difficult and uneconomical. It is very hard to find a speaker with consistent timbre and excellent multiple language skills.

Generally, code-switching speech synthesis methods boil down to two aspects. One aspect is the front-end text process. An ideal way to build a code-switching speech synthesis system is to use a bilingual speech database recorded by the same speaker [7]. Different front-end text process module for each language is deployed, which processes different languages separately. This method needs a large bilingual speech database and could not share knowledge between different languages. To integrate different languages into unified form, some researches transform the native script into another unified language. And the acoustic model is trained based the unified phonemes, which is similar to the monolingual model training [8, 9]. Based on the combined phoneme set, some researches further use two monolingual speech databases to train the code-switching model [10]. Except for the above unified phonemes based on the text analysis, an end-to-end multilingual speech recognition model called Audio-to-Byte [11] transforms each text character into the UTF-8 variable length byte based on the audio.

Another aspect for code-switching speech synthesis is the model structure. In the traditional pipeline speech synthesis framework [12], early researches choose hidden Markov model (HMM) as the code-switching acoustic model [13]. Context dependent HMM states are shared to build the Mandarin and English code-switching system [14–16]. Besides, voice conversion models are used to synthesize mixed-lingual speech [17, 18]. For the deep neural network (DNN) based speech synthesis, most of researches apply speaker and language factorization methods [19–21]. And Kullback-Leibler divergence (KLD) is used to build a code-switching speech synthesis system [22]. Nowadays, the end-to-end speech synthesis has become the current mainstream technology, which could build a high-quality code-switching system [23]. Due to the flexibility of all neural sequence-to-sequence models, learning multilingual models via conditioning on style token is straightforward [24], which could handle the code-switching phenomenon more flexibly and efficiently.

In this paper, we look into Mandarin and English code-switching end-to-end speech synthesis based on a multi-speaker bilingual speech database. Both windowing technique and style token modeling are improved for the code-switching speech synthesis system. First, it is important for the attention based encoder-decoder model to pay attention to the proper context information for each decoder time step, especially in the circumstance of context language switching. The previous different context could interfere with the current language speech generation. To improve the consistency of speaking style in bilingual situation, compared with the conventional windowing techniques that use fixed constraints, the dynamic attention reweighting soft windowing mechanism is proposed to ensure the smooth transition of code-switching. Second, to compensate the shortage of mixed-language training data, both the Mandarin and English monolingual data are added into multi-speaker bilingual speech database. The international phonetic alphabet (IPA) is used as the unified phonemes in our system, where the language dependent style token is proposed to model the discrepancy between different languages. The attention gating is proposed to adjust style token dynamically based on the language and the attended context information. Experimental results show that proposed methods lead to an improvement on intelligibility, naturalness and similarity.

http://dx.doi.org/10.21437/Interspeech.2020-1754

2. Method

Fig.1 shows the architecture of the Tacotron based code-switching end-to-end framework. Tacotron as proposed in [1] does not include explicit modeling of speaker identity in the mixed-language situation; however, due to the flexibility of all neural sequence-to-sequence models, learning multi-speaker code-switching models via conditioning on style token is straightforward. We deploy the LPCNet neural vocoder [25], which significantly improve the efficiency of speech synthesis and remains high quality. And the acoustic seq-to-seq model consists of three parts:

- Encoder mainly processes text information. Both the Mandarin and English texts are transformed into the unified IPAs, which could let the models have the shared phonemes as much as possible. The tones of Mandarin and stress of English are also included in the input text sequence.
- The attention mechanism connects the encoder and decoder. The soft windowing mechanism is based on the reweighted attention results, which aims to ignore the previous phoneme information from different languages and improve the naturalness of synthetic speech in the language switching points.
- Decoder generates the acoustic features conditioned on context and style token. For each decoder time step i, the attention gating generates shift embedding $E^{(i)}_{shift}$ base on the phoneme, language and speaker embeddings. After the shifting procedure, the language dependent shifted style token $E^{(i)}_{ST}$ would be fed into the decoder. The introduction of language dependent style token could make full use of multi-speaker monolingual speech database and limited mono-speaker mixed-languages speech database. And the decoder generates acoustic features conditioned on style token to ensure high similarity and consistency of synthetic speech.

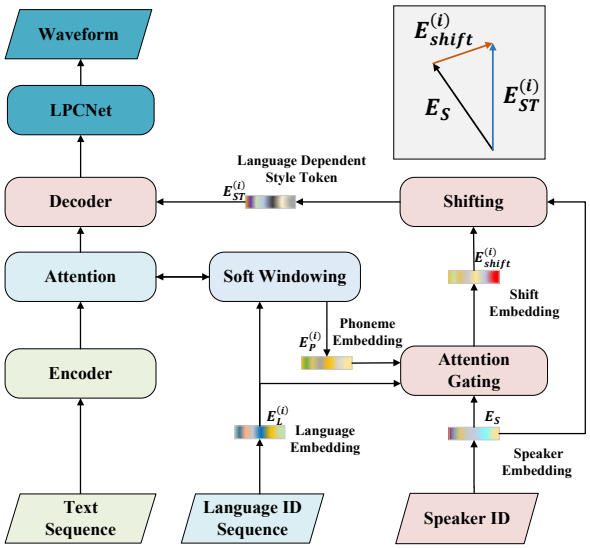

Figure 1: *System architecture of the Tacotron based code-switching end-to-end framework. The attention gating generates shift embedding $E^{(i)}_{shift}$ based on context and language information. The shifting procedure generates the language dependent style token $E^{(i)}_{ST}$. The attention weights are adjusted dynamically by the soft windowing mechanism, where the language switching flag is the key influencing factor.*

2.1. Dynamic Soft Windowing Mechanism

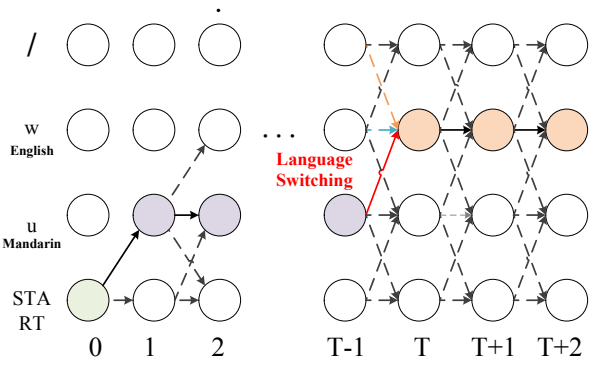

Figure 2: *Colored circles represent a possible alignment path. The red, blue and orange arrows represent move forward, remain motionless and move backward respectively.*

The dynamic soft windowing mechanism is designed to choose the scope of encoder's output sequence dynamically. To be more concrete, we expect the attention mechanism could consider the situation of language switching, where the attention mechanism could ignore previous phoneme information from different language. Unlike traditional window technology using fixed window width [26] or directly set the weights of previous phoneme information to zeros by hard way, inspired by the forward attention [27] and our previous work [28], we propose a reweighting attention mechanism to realize the dynamic soft windowing.

The alignment path in the attention mechanism indicates the mapping relation between text information and corresponding acoustic features. We assume that the alignment paths do not move strictly monotonically, which means the attended phone should move forward to the following one, remain motionless or move backward to the previous one for each decoder time step. The encoder first processes the input text sequence $X = (X_1, X_2, ..., X_U)$ to produce a sequence of hidden representations $x = (x_1, x_2, ..., x_U)$. Given the query q_t, let $\beta_{(t,n)}$ denote alignment results from local sensitive attention [1] for the index n of x at the time step t. The variable $\alpha_{(t,n)}$ is defined as the new alignment results reweighted from $\beta_{(t,n)}$. The $\alpha_{(t,n)}$ can be calculated recursively from $\alpha_{(t-1,n)}$, $\alpha_{(t-1,n-1)}$ and $\alpha_{(t-1,n+1)}$ as

$$\alpha_{(t,n)} = (\gamma^{(0)}_{t-1} \cdot \alpha_{(t-1,n)} + \gamma^{(1)}_{t-1} \cdot \alpha_{(t-1,n-1)} + \gamma^{(2)}_{t-1} \cdot \alpha_{(t-1,n+1)}) \quad (1)$$

where $\gamma^{(0)}_{t-1}$, $\gamma^{(1)}_{t-1}$, $\gamma^{(2)}_{t-1}$ are the coefficients generated by the DNN. The DNN has two hidden layers and a sigmoid output layer, which predicts the probability of the attended phoneme's next move (remain motionless: $\gamma^{(0)}_{t-1}$, move forward: $\gamma^{(1)}_{t-1}$, move backward: $\gamma^{(2)}_{t-1}$). Then we define

$$\hat{\alpha}_{(t,n)} = \alpha_t(n) / \sum_n \alpha_{(t,n)} \quad (2)$$

to normalize forward variable $\alpha_{(t,n)}$. The reweighted context vector can be computed as

$$c_t = \sum_n^N \hat{\alpha}_{(t,n)} x_n \quad (3)$$

The inputs of DNN contains c_t and q_t and the language switching flag $F_S^{(i)}$. The language switching flag is defined by the attended phoneme $X_{P^{(i)}}$, where the index of current attended phoneme $P^{(i)}$ is defined by the following equation:

$$P^{(i)} = \arg \max_{1 \le j \le U} \left(\beta_{(i,j)} \right) \tag{4}$$

The language switching flag $F_S^{(i)}$ remains at 0 and only would be set to 1 while the language of attended phoneme $X_{P^{(i)}}$ is different from previous one.

Our key insight is that the attention mechanism could be immune to the effect of previous phoneme from different languages when the language switching happens. As shown in the Fig.2, the probability of move forward (red arrow in the figure) should be larger in the code-switching situation. The increasing of $\gamma_{t-1}^{(1)}$ could let the attention weights $\alpha_{(t)}$ focus on the current attended phoneme more.

2.2. Language Dependent Style Token

The language dependent style token is generated based on context and speaker information for each decoder time step. As shown in the Fig.1, the attention gating network computes the local shift embedding by learning a non-linear combination among phoneme embedding, language embedding and speaker embedding. Our key insight is that for each decoder time step depending on the attended context from different languages, the style token used to guide the decoder may differ. For instance, the acoustic features distribution of an IPA phoneme may change in the circumstance of different languages. To handle these dynamic dependencies, the gating mechanism is proposed to control the importance of each embedding.

The inputs of attention gating is defined as followed: The current attended phoneme embedding $E_P^{(i)}$ could be built by nonlinear transform according to the current phoneme $X_{P^{(i)}}$. Besides, the speaker embedding E_S and the language embedding $E_L^{(i)}$ could also be built according to input speaker ID and language ID sequence. All the above embeddings are initialized with Glorot [29] initialization.

The phoneme gate $G_P^{(i)}$ and language gate $G_L^{(i)}$ are defined by the following equations:

$$G_P^{(i)} = \sigma \left(w_P \left[E_P^{(i)}; E_S \right] + b_P \right) \tag{5}$$

$$G_L^{(i)} = \sigma \left(w_L \left[E_L^{(i)}; E_S \right] + b_L \right) \tag{6}$$

where $[;]$ denotes the operation of vector concatenation, w_P and w_L are weight vectors, b_P and b_L are biases, $\sigma(\cdot)$ represents sigmoid function.

Then the local context and language dependent shift embedding $E_{shiift}^{(i)}$ is calculated by the following equation:

$$E_{shiift}^{(i)} = G_P^{(i)} \left(w_{hp} E_P^{(i)} \right) + G_L^{(i)} \left(w_{hl} E_L^{(i)} \right) + b_h^{(i)} \tag{7}$$

where w_{hp} and w_{hl} are weight vectors, $b_h^{(i)}$ is bias vector.

After the shifting procedure, the final language dependent style token $E_{ST}^{(i)}$ is calculated by the following equation:

$$E_{ST}^{(i)} = E_S + \alpha E_{shift}^{(i)} \tag{8}$$

$$\alpha = \min \left(\beta \|\mathbf{E}_S\|_2 / \left\| \mathbf{E}_{shift}^{(i)} \right\|_2, 1 \right) \tag{9}$$

where β is the hyper-parameter. In order to avoid the magnitude of the shifted embedding $E_{shift}^{(i)}$ is too large compared with the speaker embedding E_S, the scaling factor α is designed to constrain the magnitude of the shift embedding.

3. Experimental Setup

The mixed-language Mandarin and English database we used contains one female speaker with about 12 hours, including 2 hours mixed-language corpus. To compensate the shortage of training data, we use the Blizzard Challenge 2018 dataset and our own internal dataset to extend the training data. Our internal dataset consists of 25 different professional Mandarin speakers with about 200 hours. The Blizzard Challenge dataset is an estimated 5 hours of speech from one native English speaker. All the wav files are sampled at 16kHz. In this work, we limit the input of the synthesis to 32 features: The 30-dim Bark-scale [30] cepstral coefficients, and 2 pitch parameters (period, correlation) are extracted directly from recorded speech samples. The input text is processed by our G2P frontend and transformed to the IPA phoneme sequences, which also include tone information of Mandarin and stress information of English.

For the Tacotron training, we set output layer reduction factor $r = 2$. We use the Adam optimizer with adaptative learning rate decay, which starts from 0.0001 and decays as introduced in our previous research [28]. The training batch size is 16, where all sequences are padded to a max length.

For the LPCNet training, the network is trained for 120 epochs, with a batch size of 64, each sequence consisting of 15 10-ms frames. We use the AMSGrad [31] optimization method (Adam variant) with a step size $\alpha = \alpha_0/(1 + \delta \cdot b)$ where $\alpha_0 = 0.001$, $\delta = 5 \times 10^{-5}$, and b is the batch number.

The models on which we conduct experiments include:

- **LI-Base:** The language independent Tacotron2 baseline model is trained by only mixed-language mono-speaker data but without extra language embedding [1].
- **LD-Base:** The language dependent Tacotron2 baseline model is trained by only mixed-language mono-speaker data and use one-hot to distinguish different languages and generate language embedding through nonlinear transformer [24].
- **P-*-*:** Our proposed language dependent method is denoted as P. To do ablation studies, we make several models. DSW and HW are short for proposed dynamic soft windowing mechanism and baseline hard windowing technology respectively. SST and FE is short for proposed shifted style token method and baseline separate fixed speaker embedding and language embedding method respectively. Besides, the using of the extended data is one of key variables. Therefore, NE is denoted as no extended data is used in the experiments. For instance, P-DSW-SST is our final proposed method.

We evaluate the performance of our models in terms of intelligibility, naturalness and similarity. We focus on evaluation based on the mixed-language speaker. The test sets are about 500 utterances, where are 400 utterances are mixed-language. To evaluate intelligibility, a subset for about 50 utterances is selected by sorting high frequency unacceptable errors based on baseline evaluations. 30 listeners conducted crowd-sourcing ABX preference tests and MOS tests. In each experimental group, 30 parallel sentences are selected randomly from test subset.

4. Evaluation and Discussion

4.1. Intelligibility Evaluation

Intelligibility tests are performed with metrics of case level unintelligible rate to evaluate the robustness of models. If the synthetic speech utterance contains the unacceptable errors such as skipping, repeating and mispronunciation, this unintelligible utterance would be counted. The utterance level intelligibility results are shown in the Fig.3. We could observe that the proposed P-DSW-SST method decreases the unintelligible rate about 60%, in which the extended multi-speaker database plays an important role in improving the robustness of the system. Besides, by comparing the unintelligible rate of P-DSW-SST/FE, the language dependent style token could further improve the robustness. By analyzing the synthetic speech, we find that this further improvement mainly involves phoneme related mispronunciations. It can be interpreted that the combined language dependent style token could guide the decoder generate more accurate acoustic features, which could avoid some disturbance from the noise in the separate fixed language and speaker representations.

Besides, the function of language switching flag is also evaluated based on the attention alignments results. As shown in the Fig.4, two P-DSW-NE systems with/without language switching flag are compared. The red circles in figures is the language switching points. We can observe the introduction of language switching flag could the guide the attention mechanism better generation of alignments.

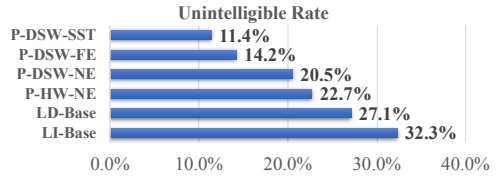

Figure 3: *Utterance level intelligibility results.*

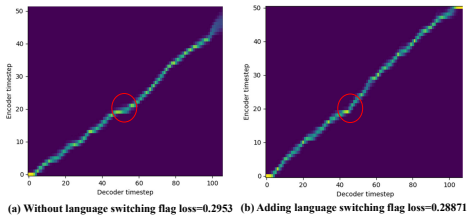

Figure 4: *Attention alignments and loss results with the same text on a test utterance in P-DSW/HW-NE systems.*

4.2. Naturalness and Similarity Evaluation

In this part, we first evaluate the naturalness of the synthetic speech from different models. The naturalness of language switching is one of key assessments. The ABX test results on the naturalness is shown in the Tab.1. By observing the preference scores, the proposed P-DSW-SST system achieve better performance than the baseline systems. Besides, the dynamic soft windowing mechanism method makes more contributions on naturalness improvement than the language dependent style token method according to the ablation results.

The similarity of the synthetic speech is one of key measures in the code-switching speech synthesis task. The MOS results on similarity of synthetic speech is illustrated in the Fig.5. The similarity consistency in different languages is evaluated. We can observe that the proposed P-DSW-FE/SST method could increase the similarity MOS for about 0.5 point, which also shows the effectiveness of the extended database. And the combined language dependent style token can achieve better performance than the separate language and speaker embeddings. We infer that the language and speaker embeddings are related. The joint shifting modeling strategy reduces redundancy and noise. Besides, the performance degradation in system P-HW-NE could also be observed. A possible explanation is that directly setting the weights of previous phoneme information to zeros could isolate the two different languages speech generation and decrease the similarity consistency.

Table 1: *Preference scores on naturalness of synthetic speech.*

System A	Scores A (%)	Scores Neutral (%)	Scores B (%)	System B
P-DSW-SST	65.83	8.69	25.48	LI-Base
	61.15	5.84	33.01	LD-Base
	55.94	7.52	36.54	P-HW-NE
	48.25	19.35	32.40	P-DSW-NE
	43.76	16.46	39.78	P-DSW-FE

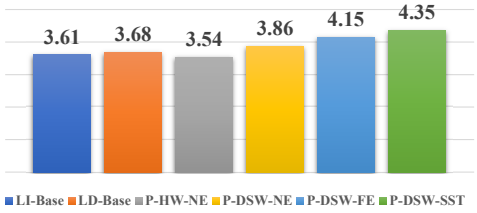

Figure 5: *MOS results on similarity of synthetic speech.*

5. Conclusions

In this paper, we propose a dynamic soft windowing mechanism and language dependent style token for code-switching end-to-end speech synthesis system. The designed language switching flag and attention reweighting mechanism could ensure the smooth transition of code-switching and improve the consistency of speaking style in bilingual situation. The language dependent style token is adjusted dynamically based on the language and the attended context information and improve the similarity by adding multi-speaker monolingual training data. Experimental results demonstrate that both methods improve intelligibility, naturalness and similarity of synthetic speech.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0820602), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379) and the Major Program for the National Social Science Fund of China (13&ZD189). This work is also supported by the CCF-Tencent Open Research Fund.

7. References

[1] 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 (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *ICLR*, 2017.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv: Computation and Language*, 2018.

[4] 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," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[5] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 3918–3926.

[6] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3422–3428. [Online]. Available: https://www.aclweb.org/anthology/L16-1546

[7] C. Traber, K. Huber, K. Nedir, B. Pfister, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*, 1999.

[8] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text." in *The International Conference on Language Resources and Evaluation*, 2016.

[9] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. W. Black, "Experiments with cross-lingual systems for synthesis of code-mixed text," in *9th ISCA Speech Synthesis Workshop*, 2016.

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

[11] 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," 2018.

[12] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[13] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *IEEE International Conference on Acoustics*, 2005.

[14] L. Hui, Q. Yao, and K. S. Frank, "An hmm-based bilingual (mandarin-english) tts," in *Sixth ISCA Workshop on Speech Synthesis*, 2007.

[15] Q. Yao, L. Hui, 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*, vol. 17, no. 6, pp. 1231–1239, 2009.

[16] Y. Qian, H. Cao, and F. K. Soong, "Hmm-based mixed-language (mandarin-english) speech synthesis," in *6th International Symposium on Chinese Spoken Language Processing*, 2008, pp. 1–4.

[17] J. He, Y. Qian, F. K. Soong, and S. Zhao, "Turning a monolingual speaker into multilingual for a mixed-language tts." in *Sixth ISCA Workshop on Speech Synthesis*, 2012, pp. 963–966.

[18] B. Ramani, M. P. Actlin Jeeva, P. Vijayalakshmi, and T. Nagarajan, "Voice conversion-based multilingual to polyglot speech synthesizer for indian languages," in *Tencon IEEE Region 10 Conference*, 2013.

[19] 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*, vol. 20, no. 6, pp. 1713–1724, 2012.

[20] 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.

[21] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis," *Interspeech 2016*, pp. 2468–2472, 2016.

[22] F. 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)*, 2016.

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

[24] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*, 2018, pp. 4693–4702.

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

[26] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Computer Science*, vol. 10, no. 4, pp. 429–439, 2015.

[27] J. X. Zhang, Z. H. Ling, and L. R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[28] R. Fu, J. Tao, Z. Wen, J. Yi, and T. Wang, "Focusing on attention: Prosody transfer and adaptative optimization strategy for multi-speaker end-to-end speech synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6709–6713.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[30] E. A. Strickland, "An introduction to the psychology of hearing (6th edition)," *Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2898–2899, 2014.

[31] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *ICLR 2018-International Conference on Learning Representations*, 2018.