# Controllable Emphatic Speech Synthesis based on Forward Attention for Expressive Speech Synthesis

*Liangqi Liu*[1,2], *Jiankun Hu*[1,2], *Zhiyong Wu*[1,2,3,*], *Song Yang*[4], *Songfan Yang*[4], *Jia Jia*[1,2], *Helen Meng*[1,3]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
[2]Beijing National Research Centre for Information Science and Technology (BNRist),
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[4] AI Lab, TAL Education Group, Beijing, China

{llq17,hujk17}@mails.tsinghua.edu.cn, {zywu,hmmeng}@se.cuhk.edu.hk,
{yangsong1,yangsongfan}@100tal.com, jjia@tsinghua.edu.cn

## Abstract

In speech interaction scenarios, speech emphasis is essential for expressing the underlying intention and attitude. Recently, end-to-end emphatic speech synthesis greatly improves the naturalness of synthetic speech, but also brings new problems: 1) lack of interpretability for how emphatic codes affect the model; 2) no separate control of emphasis on duration and on intonation and energy. We propose a novel way to build an interpretable and controllable emphatic speech synthesis framework based on forward attention. Firstly, we explicitly model the local variation of speaking rate for emphasized words and neutral words with modified forward attention to manifest emphasized words in terms of duration. The 2-layers LSTM in decoder is further divided into attention-RNN and decoder-RNN to disentangle the influence of emphasis on duration and on intonation and energy. The emphasis information is injected into decoder-RNN for highlighting emphasized words in the aspects of intonation and energy. Experimental results have shown that our model can not only provide separate control of emphasis on duration and on intonation and energy, but also generate more robust and prominent emphatic speech with high quality and naturalness.

**Index Terms**: expressive speech synthesis, emphatic speech synthesis, forward attention

## 1. Introduction

Speech emphasis plays an important role in distinguishing the focus of the utterance from the rest and conveying the underlying intention and attitude [1]. In speech interaction scenarios, synthesizing emphasis helps computer systems to express semantics and emotions more accurately and further enhance user experience, is thus attracting increasing interest.

To synthesize emphasis, various techniques are employed. For example, hidden Markov model (HMM) based speech synthesis models are employed to generate emphatic speech by constructing decision tree (DT) with emphasis-related questions [2] [3] [4]. DNN based speech synthesis models are also widely used for emphatic speech synthesis [5], and can efficiently handle the emphatic data sparsity problem in HMM-based models with shared parameters by augmenting the network input using emphasis-specific codes. With the rapid progress of end-to-end (E2E) text-to-speech (TTS) models [6] [7], nowadays E2E based emphatic speech models [8] can achieve better naturalness in synthetic speech than traditional DNN based models, which is close to human speech recording.

The flexibility and controllability of speech synthesis systems are important factors, in addition to the quality of the synthetic speech. DNN-based models commonly use emphatic-specific codes to control duration model and acoustic model respectively [5]. The integrated E2E models can generate speech with very high voice quality and naturalness [8] [9] by simplifying the synthesis stages and replacing duration model with attention mechanism, but brings new problems: 1) lack of interpretability for how emphatic codes affect the model; 2) no separate control of emphasis on duration and on intonation and energy.

Emphasized words usually manifest themselves by a slower speaking rate in the word, an increased pitch in the intonation, a higher energy in the word, or a combination of these features [10]. [11] proposed to use forward attention to decides whether to move forward or stay at each decoder time step, and further control the speed of synthesized speech. Inspired by the fact that forward attention can control the global speaking rate of synthetic speech, we propose a novel way to control the local variation of speaking rate for emphasized words and neutral words with modified forward attention. As to highlighting emphasized words in terms of intonation and energy, we divide the decoder LSTM layers into attention-RNN and decoder-RNN to disentangle the influence of emphasis on duration and on intonation (F0) and energy. Emphasis information is injected into decoder-RNN to model the emphasis characteristic in intonation and energy.

The main contributions of this work can be summarized as:

1) using modified forward attention to explicitly control the local variation of speaking rate (duration) for emphasized words and neutral words.
2) dividing the decoder LSTM layers into attention-RNN and decoder-RNN to independently model the emphasis characteristic in intonation and energy.
3) providing interpretability and separate control of emphasis on duration and on intonation and energy.

[*] Corresponding author

By disentangling the acoustic correlates into duration and pitch, it provides more controllability and flexibility in modeling the acoustic realizations of emphasis, which leads to performance improvement in emphatic speech synthesis.

## 2. Forward Attention

Given an input sequence $x = [x_1, x_2, ..., x_N]$ with length $N$, the encoder first processes $x$ into a sequence of hidden representations $h = [h_1, h_2, ..., h_N]$, then decoder generates each output $o_t$ conditioned on a distinct context vector $c_t$. The context vector is computed by focusing on the relevant elements of $h$:

$$c_t = \sum\nolimits_{n=1}^{N} \alpha_t(n) h_n \tag{1}$$

The weight $\alpha_t$ is computed by scoring each element in $h$ separately:

$$\alpha_t = Attend(q_t, h) \tag{2}$$

where $Attend()$ is additive attention mechanism [6] [12].

Forward attention is motivated by the nature of monotonic alignment between phone sequences and acoustic sequences. To achieve monotonic alignment, [11] employs a forward algorithm to modify attention probabilities at each time step:

$$\hat{\alpha}'_t(n) = \Big(\hat{\alpha}_{t-1}(n) + \hat{\alpha}_{t-1}(n-1)\Big)\alpha_t(n) \tag{3}$$

In this way, the attended phoneme at time $t$ can only come from the attended phoneme at time $t-1$ either staying or moving forward to the next one, which thus guarantees monotonic alignment. $\hat{\alpha}'_t(n)$ is normalized to $\hat{\alpha}_t(n)$ as sum weight to compute the context vector $c_t$:

$$\hat{\alpha}_t(n) = \hat{\alpha}'_t(n) / \sum\nolimits_{m=1}^{N} \hat{\alpha}'_t(m) \tag{4}$$

$$c_t = \sum\nolimits_{n=1}^{N} \hat{\alpha}_t(n) h_n \tag{5}$$

Equation (3) implies the assumption of equal probability between staying and moving forward during alignment. Actually, such transition probability is related to the current context. To incorporate such contextual information, the forward attention with transition agent is further proposed, in which a scalar $\mu_t \in (0, 1)$ is introduced to indicate the probability that the attended phone should move forward to the next one at the $t$-th decoder time step.

$$\hat{\alpha}'_t(n) = \Big((1-\mu_{t-1})\hat{\alpha}_{t-1}(n) + \mu_{t-1}\hat{\alpha}_{t-1}(n-1)\Big)\alpha_t(n) \tag{6}$$

The computation of $\mu_t$ considers the influence of the context at current decoder time step $c_t$, the decoder output at previous time step $o_{t-1}$ and the query at current time step $q_t$:

$$\mu_t = DNN(c_t, o_{t-1}, q_t) \tag{7}$$

During generation, it is easy to control the global speed of synthesized speech by adding positive or negative bias to $\mu_t$. The complete algorithm for forward attention with transition agent is described in Algorithm 1.

## 3. Methodology

### 3.1. Model architecture

The overall architecture of the proposed emphatic speech synthesis system is illustrated in Figure 1.

**Algorithm 1** Forward Attention with Transition Agent [11]

```
Initialize:
    $\hat{\alpha}_0(1) \leftarrow 1$
    $\hat{\alpha}_0(n) \leftarrow 0, n = 2, ..., N$
    $\mu_0 \leftarrow 0.5$
for $t = 1$ to $T$ do
    $\alpha_t \leftarrow Attend(q_t, h)$
    $\hat{\alpha}'_t(n) \leftarrow \Big((1-\mu_{t-1})\hat{\alpha}_{t-1}(n) + \mu_{t-1}\hat{\alpha}_{t-1}(n-1)\Big)\alpha_t(n)$
    $\hat{\alpha}_t(n) \leftarrow \hat{\alpha}'_t(n) / \sum_{m=1}^{N} \hat{\alpha}'_t(m)$
    $c_t \leftarrow \sum_{n=1}^{N} \hat{\alpha}_t(n) h_n$
    $\mu_t \leftarrow DNN(c_t, o_{t-1}, q_t)$
end for
```

We map the text to a sequence of phonemes for faster convergence and better pronunciation of rare words. The encoder first converts a phoneme sequence into a hidden feature representation. To ensure the voice quality of synthetic speech and transfer emphasis characteristic between different speakers, we extend the framework of Tacotron2 [7] following a scheme similar to [13] [14]. A learned 64-dimensional vector for the target speaker (speaker embedding) is concatenated with the encoder output at each time step.

For phonemes in the phoneme sequence, their emphatic codes (1 or 0 indicating if the corresponding phoneme is from emphasized or neutral word) compose the emphatic code sequence. Emphasis embedding is a 64-dimensional embedded vector of the emphatic codes. We replace the location-sensitive attention in Tacotron2 with the modified forward attention, and inject emphasis embedding to the modified forward attention to control the local variation of speaking rate between emphasized words and neutral words, which is significant for expressing emphasis.

Furthermore, the 2-layers LSTM in the original Tacotron2 decoder is divided into attention-RNN and decoder-RNN. Attention-RNN produces the attention query $q_t$ at each decoder time step, and the augmented context at current time $\hat{c}_t$ (containing emphasis information) generated by the forward attention module is injected into decoder-RNN to predict mel-spectrogram, thus capturing the emphasis characteristics in intonation and energy.

### 3.2. Duration control

Researches have shown that duration is the most important acoustic characteristic to distinguish between emphasized words and neutral words [15] [16]. We input emphasis information to modified forward attention to control the local speaking rate to highlight the emphasized words in terms of duration.

First, we compute emphasis context vector $z_t$ at time $t$ as the weighted sum of emphasis embeddings $e = [e_1, e_2, ..., e_N]$ similar to computing $c_t$:

$$z_t = \sum\nolimits_{n=1}^{N} \hat{\alpha}_t(n) e_n \tag{8}$$

Then, we modify forward attention by considering the influence of the emphasis context at current decoder time step $z_t$ on $\mu_t$:

$$\mu_t = DNN(c_t, o_{t-1}, q_t, z_t) \tag{9}$$

therefore we can explicitly control the local variation of speaking rate (duration) for emphasized words and neutral words.

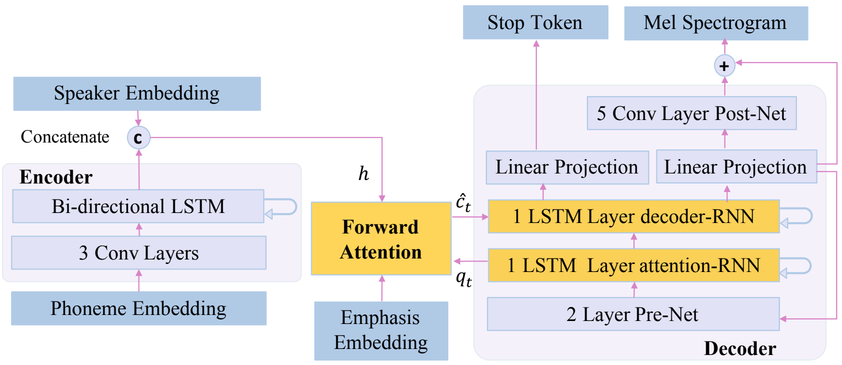

Figure 1: *The architecture of the proposed model for emphatic speech synthesis.*

### 3.3. Intonation and energy control

Different from Tacotron2 [7], the 2-layers LSTM in the decoder is divided into attention-RNN (1-layer LSTM) and decoder-RNN (1-layer LSTM) to decouple the influence of emphasis on duration and on intonation and energy. Attention-RNN generates queries at each decoder time step as input of forward attention. We concatenate context vector $c_t$ and emphasis context vector $z_t$ at current time step as augmented context vector $\hat{c}_t$:

$$\hat{c}_t = Concat(c_t, z_t) \tag{10}$$

We inject augmented context vector $\hat{c}_t$ (similar to frame-aligned input features) to decoder-RNN (acting as the acoustic model of a traditional TTS system) for controlling the prominence of emphasis in intonation and energy.

### 3.4. Emphasis strength control

The emphasis context vector at current time step $z_t$ indicates the probability of the attended phoneme at current time step coming from an emphasized word, but can't represent the strength of emphasis. During synthesis, a linear interpolation mechanism is adopted to control the strength of emphasis:

$$z_t^d = \gamma^d z_t + (1 - \gamma^d) e^{neu} \tag{11}$$

$$z_t^a = \gamma^a z_t + (1 - \gamma^a) e^{neu} \tag{12}$$

where $e^{neu}$ is the emphasis embedding corresponding to neutral emphatic code, scalars $\gamma^d \in [0, 1]$ and $\gamma^a \in [0, 1]$ are hyperparameters for controlling the emphasis strength in terms of duration, and intonation and energy, respectively. The larger value of $\gamma^d$, the longer the duration of the emphasized words is; the larger value of $\gamma^a$, the more prominent intonation and energy are.

Then $z_t^d$ is employed to affect the decision of whether to move forward or stay at each decoder time step by updating $\mu_t$:

$$\mu_t = DNN(c_t, y_{t-1}, q_t, z_t^d) \tag{13}$$

And $z_t^a$ and context vector $c_t$ of current time step are concatenated as augmented context vector $\hat{c}_t$:

$$\hat{c}_t = Concat(c_t, z_t^a) \tag{14}$$

We input augmented context vector $\hat{c}_t$ to decoder-RNN for modeling the emphasis characteristic in intonation and energy.

The complete algorithm for controllable emphatic speech synthesis is described in Algorithm 2.

**Algorithm 2** Controllable emphatic speech synthesis based on Forward Attention

```
Initialize:
    $\hat{\alpha}_0(1) \leftarrow 1$
    $\hat{\alpha}_0(n) \leftarrow 0, n = 2, ..., N$
    $\mu_0 \leftarrow 0.5$
for $t = 1$ to $T$ do
    $\alpha_t \leftarrow Attend(q_t, h)$
    $\hat{\alpha}'_t(n) \leftarrow \Big((1-\mu_{t-1})\hat{\alpha}_{t-1}(n)+\mu_{t-1}\hat{\alpha}_{t-1}(n-1)\Big)\alpha_t(n)$
    $\hat{\alpha}_t(n) \leftarrow \hat{\alpha}'_t(n)/\sum_{m=1}^{N} \hat{\alpha}'_t(m)$
    $c_t \leftarrow \sum_{n=1}^{N} \hat{\alpha}_t(n) h_n$
    $z_t \leftarrow \sum_{n=1}^{N} \hat{\alpha}_t(n) e_n$
    $z_t^d \leftarrow \gamma^d z_t + (1 - \gamma^d) e^{neu}$
    $z_t^a \leftarrow \gamma^a z_t + (1 - \gamma^a) e^{neu}$
    $\mu_t \leftarrow DNN(c_t, y_{t-1}, q_t, z_t^d)$
    $\hat{c}_t \leftarrow Concat(c_t, z_t^a)$
end for
```

## 4. Experiments

### 4.1. Experimental setup

**Dataset.** A small-scale emphatic corpus and a large-scale neutral corpus are used in our work for experiments. The large-scale neutral corpus consists of 10,000 utterances released by DataBaker [17], which has a total length of approximately 10 hours uttered by a professional native Mandarin female speaker. The small-scale emphatic corpus consists of parallel neutral and emphatic speech recordings. 500 text prompts, each of which contains one or more emphatic words at different positions, have been carefully designed to cover all kinds of pronunciation mechanisms and context characteristics of Chinese initial-finals. A professional native Mandarin female speaker has been instructed to record the emphatic speech and the parallel neutral speech according to the text prompts and the emphasis labels.

To perform emphatic speech synthesis, traditional methods require the use of large-scale corpus with emphatic speech recordings by the target speaker, which is usually difficult to obtain. In previous research, it has been proved that it is possible to learn the emphatic characteristics from the small-scale emphatic corpus and transfer it to the target speaker (with only neural speech recordings) using the multi-speaker Tacotron framework [8]. Furthermore, according to previous researches, such as cross-lingual TTS [18] and speaker transfer learning TTS

[14], the multi-speaker Tacotron framework with such design can ensure good performance when using different independent datasets for training.

The large-scale neutral corpus ensures the model to generate speech with high naturalness and stable quality. The use of emphatic corpus helps the model to learn the emphatic characteristics from the contrastive recordings of parallel neutral and emphatic speech. After training, the model can generate emphatic synthetic speech with the speaker's timbre corresponding to the large-scale neutral corpus by feeding the neutral corpus's speaker embedding. And such synthetic speech is utilized in our evaluation experiments.

**Features.** We transform text into sequences of phonemes, tones, punctuations and prosodic boundaries, all of which are represented as 512-dimensional phoneme embeddings. The use of punctuation and prosodic boundary information can effectively improve the prosody of the synthetic speech.

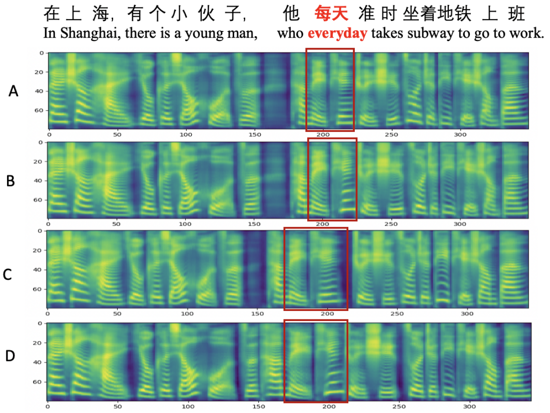

Figure 2: *Separate control of emphasis on duration and on intonation and energy. A: $\gamma^d = 0, \gamma^a = 0$, top F0 of "every day" is 338.0Hz; B: $\gamma^d = 0, \gamma^a = 1$, top F0 of "every day" is 366.3Hz; C: $\gamma^d = 1, \gamma^a = 0$, top F0 of "every day" is 263.0Hz; D: $\gamma^d = 1, \gamma^a = 1$, top F0 of "every day" is 373.2Hz.*

The speech waveforms of the two corpora are sampled at 16 kHz. Griffin-Lim [19] is used to reconstruct the waveform. Before extracting features, all waveforms are pre-emphasized with a coefficient of 0.97 as suggested by [6]. The target acoustic features were log magnitude spectrogram extracted with Hamming windowing, 50 ms frame length, 12.5ms frame shift, and 1024-point fast Fourier transform (FFT).

**Hyperparameters.** We set the reduction rate $r$=2 in all experiments. Adam optimizer [20] is used with $\beta_1$=0.9, $\beta_2$=0.999, $\epsilon$=$10^{-6}$ and fixed learning rate $10^{-3}$. All our models are trained for 50,000 global steps with a batch size of 64.

**Base model.** We extend the framework of Tacotron2 by concatenating encoder output, emphasis embedding and speaker embedding for the target speaker at each time step, which serves as the input to the decoder to predict the mel-spectrogram.

## 4.2. Experimental results and discussions

### 4.2.1. Controllability and flexibility analysis

As shown in Figure 2, the text of our example means "In Shanghai, there is a young man, who takes subway to go to work every day", in which the emphasis falls on "every day". Sub-plot A is the spectrogram specifying "every day" as neutral ($\gamma^d = 0, \gamma^a = 0$). Spectrogram in sub-plot B is produced by inputting valid emphasis information into decoder-RNN ($\gamma^d = 0, \gamma^a = 1$). Compared with A, the corresponding duration of "every day" in B is basically the same, but f0 in B is higher. Spectrogram in sub-plot C is generated by injecting valid emphasis information to update $\mu_t$ ($\gamma^d = 1, \gamma^a = 0$). Compared with A, the corresponding f0 of "every day" in C is basically the same, but duration is longer. Sub-plot D is the combination of B and C ($\gamma^d = 1, \gamma^a = 1$), f0 is higher and duration is longer for the word "every day". Note that energy is also higher but not shown in the figure. In summary, our model can control emphasis on duration and on f0 and energy respectively.

Besides, by adjusting the value of $\gamma^d$ and $\gamma^a$, we can control the strength of emphasis in terms of duration and f0, energy respectively. The larger the $\gamma^d$ is, the longer the duration of the emphasized words is; the larger the $\gamma^a$ is, the more prominent f0 and energy are. Testing samples are available at https://thuhcsi.github.io/tts/controllable-emphasis/.

Table 1: *Emphasis identification test*

| Method | Precision | Recall |
|---|---|---|
| Base Model | 80.8% | 52.5% |
| Proposed Model | 94.2% | 80.8% |

Table 2: *Naturalness test*

| Method | MOS |
|---|---|
| Base Model | 3.63(0.72) |
| Proposed Model | 3.95(0.57) |

### 4.2.2. Emphasis identification test

This experiment is designed to evaluate the perceptive accuracy of synthetic emphatic speech. 20 Mandarin native speakers with no reported listening difficulties are invited to identify all the emphasized words in 12 emphatic utterances generated by our proposed model and the base model respectively. During test, the total 24 generated utterances are randomly shuffled.

As illustrated in Table 1, our proposed model has been assessed with better perception accuracy. In particular, the recall rate has increased greatly, from 52.5% to 80.8% with a p of 0.025 in one-way ANOVA test. By disentangling the influence of emphasis on duration and on intonation and energy, our model can generate more robust and prominent emphasized words that can be perceived more easily.

### 4.2.3. Naturalness test

This experiment is designed to evaluate the naturalness and quality of generated speech in 5-point scale: 5 = Excellent (highly natural), 4 = Good (natural), 3 = Fair (clear), 2 = Poor (not clear), 1 = Bad (hard to understand). The same 20 subjects are invited to assess 12 emphatic utterances generated by our proposed model and the base model respectively.

The average mean opinion score (MOS) is presented in Table 2, one-way ANOVA test reveals our proposed model significantly outperforms the base model with a p of 0.0000015.

Our proposed model can generate emphatic speech with better quality and naturalness.

## 5. Conclusions

This paper proposes a controllable emphatic speech synthesis model based on modified forward attention. In the proposed E2E TTS framework, the duration of emphasized and neutral words can be controlled by modifying forward attention mechanism. The 2-layers LSTM in decoder is divided into attention-RNN and decoder-RNN, and emphasis information is injected into decoder-RNN for highlighting emphasized words in intonation and energy. Experimental results have shown that our model can provide separate control of emphasis on duration and on intonation and energy. Subjective experimental results confirm that our proposed approach can generate more robust and prominent emphatic speech with high quality and naturalness.

## 6. Acknowledgments

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), National Natural Science Foundation of China (61433018, 61375027), and National Social Science Foundation of China (13&ZD189).

## 7. References

[1] K. Yu, F. Mairesse, and S. J. Young, "Word-level emphasis modelling in hmm-based speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4238–4241.

[2] H. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling dnn-based speech synthesis using input codes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4905–4909.

[3] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "Hmm-based emphatic speech synthesis using unsupervised context labeling," in *ISCA Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 1849–1852.

[4] Y. Ning, Z. Wu, J. Jia, F. Meng, H. M. Meng, and L. Cai, "Hmm-based emphatic speech synthesis for corrective feedback in computer-aided pronunciation training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4934–4938.

[5] R. Li, Z. Wu, Y. Huang, J. Jia, H. Meng, and L. Cai, "Emphatic speech generation with conditioned input layer and bidirectional LSTMS for expressive speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5129–5133.

[6] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[8] M. Wang, Z. Wu, X. Wu, H. Meng, S. Kang, J. Jia, and L. Cai, "Emphatic speech synthesis and control based on characteristic transferring in end-to-end speech synthesis," in *Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 2018, pp. 1–6.

[9] Y. Zhang, S. Pan, L. He, and Z. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6945–6949.

[10] S. Shechtman and M. Mordechay, "Emphatic speech prosody prediction with deep lstm networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5119–5123.

[11] J. Zhang, Z. Ling, and L. Dai, "Forward attention in sequence- to-sequence acoustic modeling for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4789–4793.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[13] A. Gibiansky, S. Ö. Arik, G. F. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 2962–2970.

[14] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 4485–4495.

[15] J. M. Brenier, D. M. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment*. Springer, 2000, pp. 89–127.

[17] Databaker, "Open source chinese female voice database," Databaker Technology Inc., 2019, [Online]. Available: https://www.data-baker.com/open_source.html.

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

[19] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1983, pp. 804–807.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.