MULTI-SPEAKER EMOTIONAL SPEECH SYNTHESIS WITH FINE-GRAINED PROSODY MODELING

Chunhui Lu, Xue Wen, Ruolan Liu, Xiao Chen

Samsung Research China-Beijing (SRC-B)

ABSTRACT

We present an end-to-end system for multi-speaker emotional speech synthesis. In particular, our system learns emotion classes from just two speakers then generalizes these classes to other speakers from whom no emotional data was seen. We address the problem by integrating disentangled, fine-grained prosody features with global, sentence-level emotion embedding. These fine-grained features learn to represent local prosodic variations disentangled from speaker, tone and global emotion label. Compared to systems that model emotions at sentence level only, our method achieves higher ratings in naturalness and expressiveness, while retaining comparable speaker similarity ratings.

Index Terms— Multi-speaker, fine-grained, prosody modeling, emotional speech synthesis

1. INTRODUCTION

Emotional speech synthesis (ESS) aims to generate natural and expressive speech with prescribed emotion, usually one chosen from several predefined emotion classes (happy, angry, etc.). Just like humans modulating voice emotions for effective communication, ESS has the potential for improving human-compute voice interface. This is desirable in voice-enabled applications like spoken dialogue systems, spoken language translation and content creation using text-to-speech (TTS) synthesis.

While recent state-of-the-art speech synthesizers, e.g. [1, 2], take advantage of large multi-speaker datasets, obtaining emotional speech on similar scale can be challenging and expensive. One particular difficulty is that consistently producing speech with prescribed emotion needs professional training, and most of our "corpus speakers" cannot do it. In this paper we consider a more practical setup in which we have a few "emotional speakers", who are professionals providing emotional speech examples, and many "neutral speakers", from whom we have only emotionally neutral examples. More concretely, we use two emotional speakers (1 male and 1 female), and 8 neutral speakers (4 and 4). We define four emotion classes: angry, happy, sad, neutral. Our goal is to produce emotional speech in the voice of all 10 speakers.

Most previous ESS systems [3–5] use global, sentence-level emotion representation for emotion control. While this is reasonable for large, balanced training data, the same cannot be taken for granted in our setup, where most speakers are never heard with emotion. To check this out we ran a set of preliminary studies using global emotion embedding, and observed neutral speakers' synthesized emotional speech is much less natural and expressive than that of emotional speakers. In other words, the system had learned to associate emotions mostly with emotional speakers. This suggests that we look for a way to represent emotion that is detached from speaker identity.

In this paper, we augment sentence-level emotion embedding with fine-grained prosody representation. The latter captures local speech variations not fully characterized by text, speaker ID and global emotion class. Concretely, we associate each spoken phoneme with a 3-dimension latent code learned within a conditional variational autoencoder (VAE) framework. This very tight 3-wide information bottleneck encourages learning acoustically important features not already present among the conditioning inputs, therefore detaches the latent code from speaker identity. By transferring this latent code from emotional speakers, we are able to produce emotional speech of neutral speakers. Experimental results show our method achieves consistent improvements in both naturalness and expressiveness.

2. RELATED WORK

2.1. Single speaker ESS

Most ESS systems so far use emotional speech from one single speaker. [6, 7] implemented emotional statistical parametric speech synthesis (SPSS) with emotion codes. [3] introduced an end-to-end (E2E) emotional speech synthesizer based on Tacotron [8] by injecting a learned emotion embedding. [4, 5] adopted pretrained global style tokens (GSTs [9]) to represent different emotions. Some systems also considered synthesizing emotion at different strength levels [10, 11].

2.2. Multi-speaker ESS

For multi-speaker ESS, [12] investigated different combinations of speaker and emotion representations with a convolutional neural network (CNN) synthesizer. They used 10 speakers, each with examples in all emotion classes. For generalizing emotional expressions to new speakers, early SPSS researches [13, 14] tried to represent emotion as an additive factors. [15] investigated several deep neural network (DNN) architectures for emotion transplantation. In this work they used 3 emotional and 21 neutral speakers.

2.3. Fine-grained prosody modeling

Though sentence-level latent representation shows the ability to capture prosodic features from speech [9, 16], it is short of fine-grained controllability and robustness in inter-speaker transfer. Recently,

a train of researches have turned to fine-grained prosody modeling [17–20]. [17] introduced temporal structures on both speech and text sides for prosody embedding, which enabled pitch and amplitude manipulation at frame and phoneme levels. [18] pre-computed prosody-related acoustic features of phonemes, and used them to reproduce a reference prosody on synthesized speech. [19] and [20] used fine-grained VAEs to learn local latent features under different model specifications. Notably, both [17] and [19] used low-dimension latents that explicitly associated with meaningful acoustic attributes.

In this paper, we look at both the multi-speaker and fine-grained prosody modeling aspects of ESS.

3. METHOD

3.1. Baseline

Our baseline synthesizer is based on FastSpeech [21], an E2E TTS system based on self-attentional convnet taking additional phoneme duration input. We adapt FastSpeech for multi-speaker ESS as shown in Fig. 1.

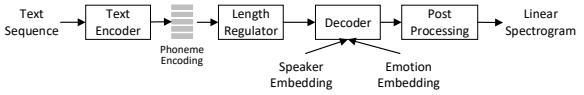


Fig. 1. Baseline architecture

The text input is a phoneme-tone-(phone_duration) sequence, where the tone label applies to tonal languages like Mandarin Chinese. For each phoneme unit, we concatenate a positional encoding to its phoneme and tone embeddings and feed the joint vector to the text encoder. The length regulator repeats each phoneme encoding by its duration in frames. The regulated text encoding has the same length as the speech. The decoder takes this regulated encoding sequence as input to predict the mel spectrogram, conditioned on speaker and emotion embeddings. We investigate two approaches [3, 5] to emotion embedding, either as a free, learnable vector which we denote as BASE-EMB, or as weighted combination of GSTs which we denote as BASE-GST.

We follow [21] to predict phoneme duration in log domain within FastSpeech, with the difference that we now predict duration conditioned on speaker and emotion. Reference durations are obtained by force alignment[1].

We use a post-processing network [8] to convert mel spectrogram to linear spectrogram and the Griffin-Lim [22] algorithm to construct audio output.

3.2. Fine-grained prosody modeling

We introduce fine-grained latents to BASE-EMB to learn disentangled local variations in emotional speech. As shown in Fig. 2, phoneme-aligned spectrogram is sent to a reference encoder, conditioned on tone, speaker and emotion information. For each phoneme, the reference encoder computes a variational posterior over its latent, from which we draw a sample and append it to the phoneme encoding before sending to the length regulator. We apply domain adversarial training [23] to further disentangle the latent code from speaker and tone, implemented with two discriminators and a gradient reversal unit [24].

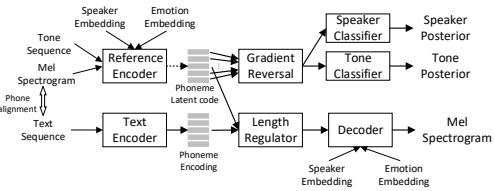


Fig. 2. Proposed architecture. Dashed lines denote sampling via reparameterization [25].

The objective function can be formulated as combining an evidence lower bound (ELBO) with a domain adversarial objective:

$$\mathcal{L} = \mathcal{L}_{ELBO} - \lambda_{adv}\mathcal{L}_{adv} \tag{1}$$

where $\mathcal{L}_{adv}$ is the sum of minimum cross-entropy losses for the two classifiers. $\mathcal{L}_{ELBO}$ is actually a β-VAE objective [26] under standard Gaussian latent prior:

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y},\mathbf{s},\mathbf{e})}[\log p(\mathbf{x}|\mathbf{y},\mathbf{z},\mathbf{s},\mathbf{e})] \\ &- \lambda_{KL}\sum\nolimits_{u=1}^{U} D_{KL}(q(\mathbf{z}_u|\mathbf{x},\mathbf{y},\mathbf{s},\mathbf{e}))\|\mathcal{N}(0,I))\end{aligned} \tag{2}$$

where $\mathbf{z}$ represents the sequence of latents and $\mathbf{z}_u$ is the latent code for the u-th phoneme. U, s, e are the number of phones, speaker embedding and emotion embedding, respectively. We will use $0<\lambda_{KL}<1$, which favors accuracy over latent space exploration.

3.3. Latent codes prediction

Once the TTS model is trained, we can use the reference encoder to collect latent mean values from emotional speech. We may then train a separate model to predict them from text, speaker and emotion class, under e.g. the mean square error (MSE) loss. This model only sees emotional speakers. For emotional synthesis of neutral speakers, we pick one emotional speaker to predict the latents, then use the target speaker for decoding.

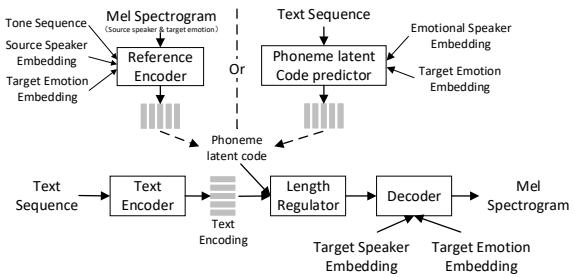


Fig. 3. The diagram during inference. Phoneme latents are extracted from reference mel spectrogram or predicted from input text.

[1]Using an open-source Kaldi ASR model: http://www.kaldi-asr.org/models/m2

Alternatively, if we are tasked with rendering an emotional utterance in a neutral target speaker's voice, the latent code may also be computed using the reference encoder, as shown in the top left part of Fig. 3. A forced alignment module is needed for this type of emotion transfer.

4. EXPERIMENTS

4.1. Setup

We conduct experiments on internal Mandarin Chinese datasets, detailed in Table 1. The two emotional speakers are professional voice actors; the others are non-professionals. All utterances come with transcriptions as phoneme-tone sequences. Recordings are sampled at 24kHz. 80-dim mel-scale and 1025-dim linear-scale spectrograms are computed from 50ms windows at 12.5ms intervals.

Table 1. Training data

Speaker Type	#speakers		#utterances per speaker			
	F	M	Angry	Happy	Neutral	Sad
Emotional	1	1	500	500	500	500
Neutral	4	4	-	-	1000	-
Total	10		1000	1000	9000	1000

Details of the FastSpeech network is same as [21]. Both speaker and emotion embeddings are set to 64-dim in all systems. We follow [5] for emotion embedding in BASE-GST. The reference encoder closely follows that in [16] and extracts a 3-dim latent code for each phoneme. Both tone and speaker classifiers use 2-layer feedforward networks with hidden layer size 256. We set λ_{KL}, λ_{ADV} to 0.01 and 0.02 respectively. Training the TTS engine took about 150k steps at batch size 32 on one P40 GPU. Synthesized speech samples can be found here[2].

4.2. Subjective evaluation

4.2.1. MOS

We first investigate mean opinion score (MOS) in naturalness, speaker similarity and the clarity of emotional expressiveness. 15 Mandarin speakers are asked to listen to the synthesized speech and rate them on a scale between 1 and 5. For similarity, target speaker's neutral speech is used as reference. Compared systems are as follows.

- GT: ground-truth speech reconstructed by Griffin-Lim.
- BASE-EMB: baseline method using learnable emotion embedding as emotion representation.
- BASE-GST: baseline method using weighted combination of GSTs to represent different emotions. The reference audio samples of each emotion are selected from the visualization cluster center and used when synthesizing target emotion.
- OURS-TRANS: proposed method with latent codes transferred from emotional speakers' emotional test speech.
- OURS-SYN: proposed method with latent codes predicted from text.

[2]https://chunhui-lu.github.io/ICASSP2021/index.html

The results are given in Tables 2 and 3. Neutral speakers (N) and emotional speakers (E) are calculated separately.

Naturalness: For angry and happy examples, we observe little difference among systems for emotional speakers, but significant improvements with our fine-grained approach for neutral speakers. This shows that explicitly modeling phoneme-level variations can indeed help generalize emotions to neutral speakers. Difference between OURS-SYN and OURS-TRANS is small, showing that the latent means are indeed predictable. For neutral examples, emotional speakers is rated 0.3 higher than neutral speaker by ground truth, indicating professional speakers sound more natural than non-professionals. The same gap is also observed in all systems tested, and we do not find notable difference between baseline and proposed systems. For sad examples, we observe low ratings across all systems (except BASE-GST). We conjecture that slow speaking rate with low F0 can give an unnatural feeling. Comparison between systems remain inconclusive.

Speaker similarity: Similarity is rated higher for neutral examples than other emotions across all systems and speakers, but the gap is particularly large for emotional speakers. This may be related to the wide prosody variations in emotional speech, and we do notice professional speakers modulate their voice traits to deliver diverse emotions. Within each speaker type (E or N), similarity ratings are similar across systems (again except BASE-GST+sad). The scores of sad examples are again notably lower. We observe that F0 differences between sad and neutral are bigger than other emotions, which may partially explain the observation.

Emotional expressiveness: For angry and happy examples of neutral speakers, we observe significant improvement in emotional expressiveness with our models over the baselines, with ratings on par with those for emotional speakers. This shows fine-grained modeling adds to the model's capability to deliver these emotions. For sad examples, the ratings are similar across systems, maybe because global traits like long duration and low F0 are already sufficient to express sadness so local variations add little to them.

4.2.2. A-B preference tests

We run an A-B preference test to probe the effect of emotional phoneme duration models. In each pair of synthesized examples, one has duration input predicted with target emotion, the other predicted with neutral emotion. Raters are asked to give their preference in terms of the naturalness. The results are given in Fig. 4.

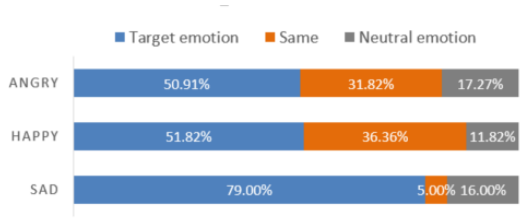


Fig. 4. AB preference results on naturalness of duration input predicted with different emotions

Table 2. MOS on naturalness and speaker similarity

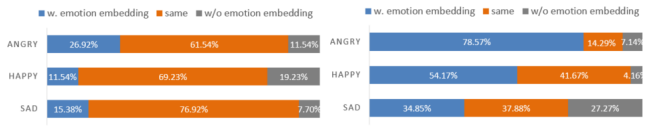
Type	Method	Naturalness				Speaker Similarity			
		Angry	Happy	Neutral	Sad	Angry	Happy	Neutral	Sad
N	GT	-	-	4.29 ± 0.23	-	-	-	4.70 ± 0.15	-
E	GT	4.86 ± 0.15	4.73 ± 0.17	4.61 ± 0.18	4.38 ± 0.25	3.51 ± 0.36	3.63 ± 0.30	4.68 ± 0.17	3.58 ± 0.20
N	BASE-EMB	3.80 ± 0.36	3.87 ± 0.34	4.02 ± 0.25	3.48 ± 0.25	3.92 ± 0.30	3.96 ± 0.24	4.42 ± 0.17	3.39 ± 0.36
	BASE-GST	3.78 ± 0.23	3.83 ± 0.23	4.00 ± 0.28	3.90 ± 0.24	3.90 ± 0.29	4.21 ± 0.21	4.48 ± 0.19	3.93 ± 0.28
	OURS-TRANS	4.43 ± 0.36	4.41 ± 0.29	-	3.74 ± 0.26	3.93 ± 0.28	4.01 ± 0.28	-	3.33 ± 0.34
	OURS-SYN	4.42 ± 0.33	4.20 ± 0.21	3.94 ± 0.28	3.65 ± 0.23	4.02 ± 0.27	4.03 ± 0.26	4.49 ± 0.17	3.43 ± 0.33
E	BASE-EMB	4.41 ± 0.31	4.34 ± 0.28	4.30 ± 0.26	3.63 ± 0.32	3.57 ± 0.34	3.61 ± 0.31	4.43 ± 0.25	3.31 ± 0.16
	BASE-GST	4.37 ± 0.26	4.29 ± 0.19	4.30 ± 0.23	3.64 ± 0.28	3.57 ± 0.20	3.58 ± 0.19	4.41 ± 0.21	3.31 ± 0.21
	OURS-SYN	4.43 ± 0.29	4.43 ± 0.29	4.26 ± 0.30	3.88 ± 0.23	3.53 ± 0.27	3.65 ± 0.31	4.43 ± 0.16	3.41 ± 0.26

Table 3. MOS on the clarity of emotional expressiveness

Type	Method	Angry	Happy	Sad
E	GT	4.92 ± 0.09	4.89 ± 0.07	4.75 ± 0.09
N	BASE-EMB	3.72 ± 0.27	3.75 ± 0.36	3.96 ± 0.29
	BASE-GST	3.70 ± 0.23	3.53 ± 0.20	3.57 ± 0.28
	OURS-TRANS	4.60 ± 0.19	4.55 ± 0.21	4.06 ± 0.22
	OURS-SYN	4.47 ± 0.28	4.31 ± 0.28	3.96 ± 0.25
E	BASE-EMB	4.54 ± 0.23	4.48 ± 0.26	4.02 ± 0.27
	BASE-GST	4.53 ± 0.23	4.43 ± 0.18	4.03 ± 0.26
	OURS-SYN	4.52 ± 0.28	4.55 ± 0.24	4.04 ± 0.33

It shows that emotional duration modeling is more effective for sad examples, which are much slower than neutral emotion. In angry and happy examples the effect is still significant but less dominating, as other non-duration prosodic traits have become more relevant in expressing these emotions.

Our method combines sentence- and phoneme-level emotion/prosody embeddings. To check the effect of global embedding in the presence of phoneme-level embeddings, we run an A-B preference test to compare our system against one trained without sentence-level embedding. Raters are asked to choose the one with higher emotional expressiveness clarity.



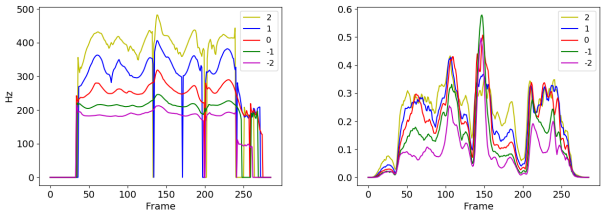
(a) Emotional speaker (b) Neutral speaker

Fig. 5. AB preference result on the clarity of emotional expressiveness of systems with and without global emotion embedding

From the results given in Fig. 5, we find that the sentence-level emotion embedding is essential for neutral speakers but not for emotional speakers. This indicates remnant entanglement between speaker and local latents, and may be related to the model not seeing emotional local latents for neutral speakers during training.

4.3. Controllability

To find out if our fine-grained system provides controllability via latent codes, for each of the three dimensions, we set the value of this dimension to one of $\{-2, -1, 0, 1, 2\}$ at each time while keeping the other two at 0. Sharing the latent codes across all phonemes, we find that the 1st dimension controls F0 of synthesized speech while the 3rd dimension controls the energy. Pitch and energy trajectories with different latent values are shown in Fig. 6. This observation reproduces that of [17], i.e. it is possible to learn disentangled and acoustically meaningful latents fully unsupervised.



(a) 1st dimension - pitch (b) 3rd dimension - energy

Fig. 6. Controllability via latent codes

5. CONCLUSION

We introduce fine-grained prosody modeling into multi-speaker ESS by learning and predicting tightly-bottlenecked phoneme-level latent features, and using these latents to condition the backbone synthesizer. These latents provide a prosody-related information path parallel to the baseline path, but much less entangled with speaker identity. This design is particularly relevant in our low-emotion-resourced setting, where we try to generalize the notion of voice emotion from just two emotional speakers to many emotionless speakers. Its effectiveness is validated by our ESS experiments, particularly for happy and angry emotions.

In future work we will consider applying similar design principle to other ESS baselines. We are also interested in probing the runtime behavior of the system, especially in hard cases like the sad emotion, as well as exploring the utility of fine-grained controllability in actual use cases.

6. REFERENCES

[1] Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *NeurIPS*, 2017, pp. 2962–2970.

[2] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *ICLR*, 2018.

[3] Younggun Lee, Azam Rabiee, and Soo-Young Lee, "Emotional end-to-end neural speech synthesizer," in *Machine Learning for Audio Signal Processing (NIPS workshop)*, 2017.

[4] Peng-Fei Wu, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, Hong-Chuan Wu, and Lirong Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *APSIPA*, 2019, pp. 623–627.

[5] Ohsung Kwon, Inseon Jang, Chunghyun Ahn, and Hong-Goo Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1383–1387, 2019.

[6] Shumin An, Zhenhua Ling, and Lirong Dai, "Emotional statistical parametric speech synthesis using lstm-rnns," in *APSIPA ASC*, 2017, pp. 1613–1616.

[7] Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai, "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, 2018.

[8] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017, pp. 4006–4010.

[9] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*, 2018, pp. 5167–5176.

[10] Xiaolian Zhu, Shan Yang, Geng Yang, and Lei Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *ASRU*, 2019, pp. 192–199.

[11] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, Chunghyun Ahn, and Hong-Goo Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP*, 2020, pp. 7254–7258.

[12] Heejin Choi, Sangjun Park, Jinuk Park, and Minsoo Hahn, "Multi-speaker emotional acoustic modeling for cnn-based speech synthesis," in *ICASSP*, 2019, pp. 6950–6954.

[13] Langzhou Chen, Norbert Braunschweiler, and Mark J. F. Gales, "Speaker and expression factorization for audiobook data: Expressiveness and transplantation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 605–618, 2015.

[14] Yamato Ohtani, Yu Nasu, Masahiro Morita, and Masami Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *Interspeech*, 2015, pp. 274–278.

[15] Katsuki Inoue, Sunao Hara, Masanobu Abe, Nobukatsu Hojo, and Yusuke Ijima, "An investigation to transplant emotional expressions in dnn-based TTS synthesis," in *APSIPA*, 2017, pp. 1253–1258.

[16] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *ICML*, 2018, pp. 4700–4709.

[17] Younggun Lee and Taesu Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP*, 2019, pp. 5911–5915.

[18] Viacheslav Klimkov, Srikanth Ronanki, Jonas Rohnke, and Thomas Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in *Interspeech*, 2019, pp. 4440–4444.

[19] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP*, 2020, pp. 6264–6268.

[20] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior," in *ICASSP*, 2020, pp. 6699–6703.

[21] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," in *NeurIPS*, 2019, pp. 3165–3174.

[22] D. W. Griffin and Jae S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[23] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2017.

[24] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," in *Interspeech*, 2019, pp. 2080–2084.

[25] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[26] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.