[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. NAACL-HLT*, Minneapolis, U.S.A., pp. 4171–4186 (2019).
[12] Zen, H., Tokuda, K. and Black, A.: Statistical Parametric Speech Synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).
[13] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A.: Tacotron: Towards End-to-End Speech Synthesis, *Proc. INTERSPEECH*, Stockholm, Sweden, pp. 4006–4010 (2017).
[14] v. d. Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, Vol. abs/1609.03499 (online), available from ⟨http://arxiv.org/abs/1609.03499⟩ (2016).
[15] Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A. and Bengio, Y.: Char2Wav: End-to-End Speech Synthesis, *Proc. ICLR Workshop*, Toulon, France (2017).
[16] Ping, W., Peng, K. and Chen, J.: ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech, *Proc. ICLR*, New Orleans, U.S.A. (2019).
[17] Takamichi, S., Sonobe, R., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JSUT and JVS: Free Japanese Voice Corpora for Accelerating Speech Synthesis Research, *Acoustical Science and Technology*, Vol. 41, No. 5, pp. 761–768 (2020).
[18] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proc. ICLR*, San Diego, California, U.S.A. (2015).
[19] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Proc. NeurIPS*, Vol. 33, Virtual Conference, pp. 17022–17033 (2020).
[20] Sugiura, K., Shiga, Y., Kawai, H., Misu, T. and Hori, C.: A Cloud Robotics Approach towards Dialogue-Oriented Robot Speech, *Advanced Robotics*, Vol. 29, No. 7, pp. 449–456 (2015).
[21] Jia, Y., Zen, H., Shen, J., Zhang, Y. and Wu, Y.: PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS, *Proc. INTERSPEECH*, Brno, Czech Republic, pp. 151–155 (2021).