Phoneme-Level Speaking Rate Variation on Waveform Generation using GAN-TTS

Mayuko Okamoto
Nara Institute of
Science and Technology
Japan
okamoto.mayuko.oi1@is.naist.jp

Sakriani Sakti
Nara Institute of Science and Technology
RIKEN, Advanced Intelligence Project AIP
Japan
ssakti@is.naist.jp

Satoshi Nakamura
Nara Institute of Science and Technology
RIKEN, Advanced Intelligence Project AIP
Japan
s-nakamura@is.naist.jp

Abstract—**The development of text-to-speech synthesis (TTS) systems continues to advance, and the naturalness of their generated speech has significantly improved. But most TTS systems now learn from data using a deep learning framework and generate the output at a monotonous speaking rate. In contrast, humans vary their speaking rates and tend to slow down to emphasize words to distinguish elements of focus in an utterance. Unfortunately, recording natural speech with various speaking rates is expensive and time-consuming. This paper constructs synthetic and natural speech corpora with a variable speaking rate and analyzes the main difference in the speaking rates of natural and artificial data. We develop a generative adversarial network (GAN) based TTS that enables waveform generation with phoneme-level speaking rate variations.**

Index Terms—**Text-to-speech synthesis, generative adversarial networks, speaking rate variation**

I. Introduction

A fundamental technology that creates a machine that can communicate with humans through natural conversation by speech is a text-to-speech synthesizer (TTS), which enables computers to learn how to speak. Various TTS approaches have been developed, including a waveform unit concatenation approach [1], [2], the statistical modeling of a hidden Markov model (HMM) [3], such a deep learning framework as an end-to-end deep neural network [4], [5], a wavenet [6], and generative adversarial networks (GANs) [7].

TTS systems, which are currently used in a wide range of applications, successfully produce speech with a high degree of intelligibility. The naturalness of the generated speech has also significantly improved. Unfortunately, they cannot be regarded as natural-sounding devices. The speaking styles and speech expressions produced by the current text-to-speech systems are typically averaged over training material that was mainly collected only in reading-style speech. Therefore, the speech lacks the variety and liveliness found in natural speech. But, recording a large amount of speech utterances that cover various speaking styles and expressive ranges is time-consuming and expensive.

Several studies addressed this issue by modeling speaking style variations. Yoshimura et al. simultaneous modeled the spectrum, pitch, and duration in HMM-based speech synthesis so that their system can generate speech that resembles various speaker's voices [8]. Yamagishi et al. achieved emotional expressivity and speaking style variability in an HMM-based speech synthesis [9]. But these techniques were based on the HMM framework. For deep learning, Skerry-Ryan et al. augmented an end-to-end Tacotron with explicit prosody controls for expressive speech synthesis.

Unfortunately, only few studies have addressed the speaking rate issue. In fact, the speaking rate often significantly influences how listeners perceive speech. We may speak much quicker during an emergency or slow down for greater emphasis. A study by Manson et al. found that when the speech rate is entrained and where dyad speech rates converged from the beginning to the end of a conversation, the success rate of negotiation and cooperation will probably increase [10]. Therefore, it is critical to developing a speech synthesis system that can produce natural spoken dialogues by considering the other party, as well as be able to phrase the message with an appropriate speaking rate according to the situation.

Recently, Wang et al. proposed "global style tokens" (jointly trained within Tacotron) that can be used to control synthesis speech by varying the speed and speaking style [11]. But since its information embedding is stored globally, controlling the duration at the word and phoneme levels is challenging. Although Park et al. introduced a mechanism for phonemic-level duration within the sequence-to-sequence framework [12], their system required phoneme input instead of text, which may be too complicated for actual users. The proposed method's effectiveness was also only evaluated through simulated data with unnatural speaking rate variations. Furthermore, the approach did not consider the effect on the other party.

In this study, we first construct synthetic and natural speech corpora with variable speaking rates, analyze the main difference in their speaking rates, and develop a TTS that enables waveform generation with phoneme-level speaking rate variations. In contrast to existing works, our work is based on GAN-TTS because it can also achieve a high-quality speech synthesis using only a small amount of data. We also subjectively evaluate our system and investigate its effect on listeners when it produces speech at a constantly fast speed or when it produces speech that slows down to emphasize a certain message.

978-1-7281-2449-0/19/$31.00 © 2019 IEEE

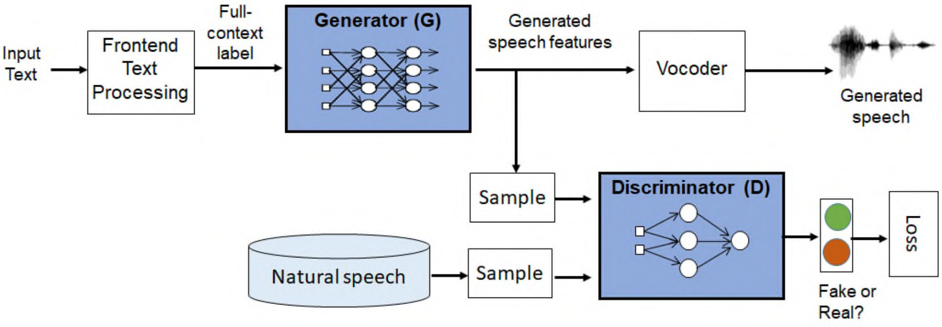


Fig. 1. Overview of GAN-TTS.

II. Speech synthesis using a generative adversarial network

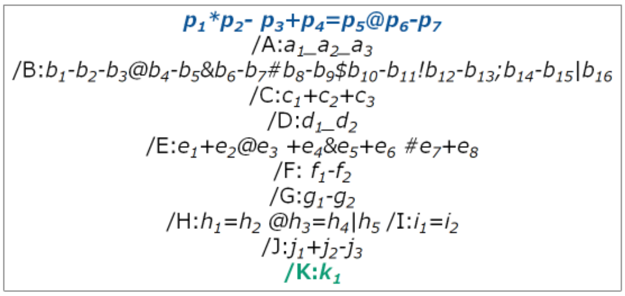
Two techniques remain widely studied in speech synthesis methods: HMM [3] and deep learning frameworks [5]. Although HMM-based speech synthesis works faster using just a small amount of training data, its sound quality is lower than DNN speech synthesis. On the other hand, such end-to-end DNN-based speech synthesis technology like Tacotron, which was rapidly developed, can easily generate acoustic speech features directly from characters. But the disadvantage of a method that uses such a DNN-based framework is that a large amount of data is required for learning. Consequently, to enable a speech synthesis system that can control the speech speed, we require a large amount of data containing various speech speeds. However, such data are often unavailable. Therefore, we utilize GAN-TTS [7], a speech synthesis that is constructed based on a generation adversarial network that can produce high-quality speech by learning even from a relatively small amount of data.

Figure 1 illustrates the GAN-TTS architecture, which consists of two types of neural networks: generator G and discriminator D. Its training procedure is employed by an adversarial process in which the two models (generator G and discriminator D) compete. In other words, the generator learns to create the speech output that causes the discriminator to misrecognize the generated result as natural speech, and the discriminator learns to accurately distinguish between natural and synthetic speech produced by generator G. Further details of GAN-TTS technology are available [7].

III. Proposed method

As described in Fig. 2, a front-end text processing block extracts the linguistic features from a given input text. Since many contextual factors (e.g., phoneme identity, word stress, etc.) might affect speech's prosodic characteristic, generating a full-context label from a given text is the most common way in standard GAN-TTS, which is also well-known in a HMM-based TTS framework.

Figure 2 shows an example of a full-context label that is comprised of the following factors:



p_1*p_2- p_3+p_4=p_5@p_6-p_7
/A:a_1_a_2_a_3
/B:b_1-b_2-b_3@b_4-b_5&b_6-b_7#b_8-b_9\$$b_{10}$-$b_{11}$!$b_{12}$-$b_{13}$;$b_{14}$-$b_{15}$|$b_{16}$
/C:c_1+c_2+c_3
/D:d_1_d_2
/E:e_1+e_2@e_3 +e_4&e_5+e_6 #e_7+e_8
/F: f_1-f_2
/G:g_1-g_2
/H:h_1=h_2 @h_3=h_4|h_5 /I:i_1=i_2
/J:j_1+j_2-j_3
/K:k_1

Fig. 2. Example of full-context label

- Phoneme level:
  - $p_1, ..., p_5$: {second preceding, preceding, current, succeeding, second succeeding} phoneme;
  - p_6, p_7: position of current phoneme in the current word (forward and backward);
- Syllable level:
  - $a_1, ..., a_3$: {type of syllable stress, number of phonemes} in the preceding syllable.
  - $b_1, ..., b_16$: {type of syllable stress, number of phonemes, position in word and phrase, number of syllables before and after} in the current syllable.
  - $c_1, ..., c_3$: {type of syllable stress, number of phonemes} in the succeeding syllable.
- Word level:
  - d_1, d_2: {part-of-speech, number of syllables} in the preceding word.
  - $e_1, ..., e_8$: {part-of-speech, number of syllables, position in phrase, number of content words before and after} in the current word.
  - f_1, f_2: {part-of-speech, number of syllables} in the succeeding word.
- Phrase level:
  - g_1, g_2: number of {syllables and words} in the preceding phrase.
  - $h_1, ..., e_5$: {number of syllables and words, utterance position, TOBI endtone} of the current phrase.
  - i_1, i_2: number of {syllables and words} in the succeeding phrase.

- Utterance level:
  - $j_1, ..., j_3$: number of {syllables, words, and phrases} in the utterance;

To achieve a GAN-TTS that controls the speaking rate variations at the phoneme level, we propose two ways to incorporate such speed information within GAN-TTS:

1) Method-1: phoneme-level incorporation
   Here we incorporate the information of the speaking rate variations within the phoneme symbols of the full-context label.
   - First, we define three discrete symbols of speech speed tags that differentiate the speaking rate: "N" for standard read speech (normal), "S" for slow, and "F" for fast speech.
   - Then we incorporate the speech speed tag within the phoneme level by directly attaching it to the phoneme label by specifically modifying $p_1, ..., p_5$ label into $p_1 + N/S/F, ..., p_5 + N/S/F$.
   - For example, the standard label of phoneme "eh" in word "$hello \rightarrow (hh, eh, l, ow)$" is "$pau * hh - eh + l = ow$". With the slow speaking rate information, it becomes "$pau * hhS - ehS + lS = owS$".
2) Method-2: utterance-level incorporation
   Here we incorporate the information of the speaking rate variations within the utterance level of the full-context label.
   - In such audio manipulation tools as SoundExchange (sox) [13], we define the percentage that modifies the original rate's speed (i.e., 75% to slow down or 125% to speed up).
   - Here we set 100 as the normal speed, below that level is slow speech, and over it is fast speech.
   - We incorporate the speech speed tag within the utterance-level by introducing new tag $K : k_1$ after $j_1, ..., j_3$. In k_1, we define the speed as a percentage.
   - For example, in the previous example of word "hello (hh eh l ow)," the phoneme label is kept the same as in standard label "$pau*hh-eh+l = ow$," but at the utterance level, we added "K : 75" to explain that we modified speaking rate 75% from the original rate.

In both methods, "pau" for the silent part ignores the speaking rate information, since there is no change in the generated acoustic features regardless of the speed information.

IV. Data construction and analysis

A. Construction of speech data with different speaking rate

Our data are based on the CMU ARCTIC database that was constructed at the Carnegie Mellon University as phonetically balanced, US English single speaker databases designed for unit selection speech synthesis research [14]. It consists of 1132 utterances in a reading speech style at a normal speed.

We performed the following procedure:

- Prepared data samples
  First, we created data samples from the CMU ARCTIC database based on the original "normal "data. After that, we modified the speaking rate artificially using sox with parameters of 0.75 (75% slowing down the original rate) and 1.25 (125% speeding up the original rate) to create "slow" and "fast" sample data. We chose these parameters in which the resulting slow and fast speech still sound natural.
- Recorded natural speech
  Next we recorded the natural speech data uttered by one female and one male. We asked them to talk as naturally as possible and simultaneously produce speech with three different speaking rates, as in the data samples. We recorded at a 44.1-kHz sampling rate and 16-bit depth. We got 6,792 utterances (two speakers and three speaking rates) that were used for analysis as well as model learning and evaluation.
- Created artificial data
  To compare and analyze the natural and artificial data with the same speaker voice, we artificially modified the speaking rate from the "normal" version of the newly recorded data. Similar to before, we used sox with parameters of 0.75 (75% slowing down the original rate) and 1.25 (125% speeding up the original rate) for "slow" and "fast" from the new data with female and male speakers.

B. Analysis

Several analyses were carried out in this study. First, we investigated the average duration of the total utterances. Fig. 3 shows the average speech length in "normal," "slow," and "fast" for female and male voices. The continues values represent the ratio between the "slow" and "fast" conditions with respect to the "normal" condition. The one from the artificial data is marked "a_slow" and "a_fast." The artificial data have 0.75 and 1.25 for the "slow" or "fast" condition in exactly the same way as we established the parameters during the data creation. Given the same reference, humans tend to produce slower speech than the artificial one in the "slow" condition and faster speech than the artificial one in the "fast"

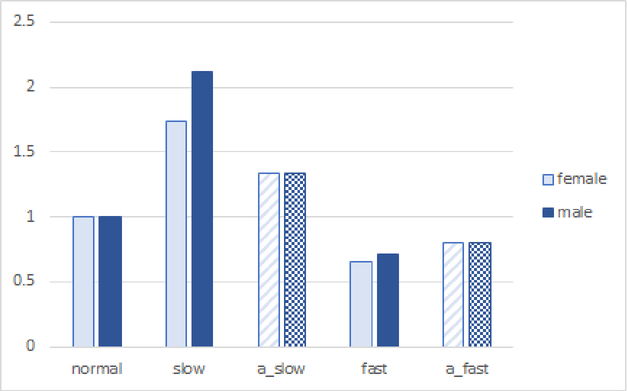


Fig. 3. The average speech length in "normal," "slow," and "fast" for female and male voices. The artificial data is marked "a_slow" and "a_fast."

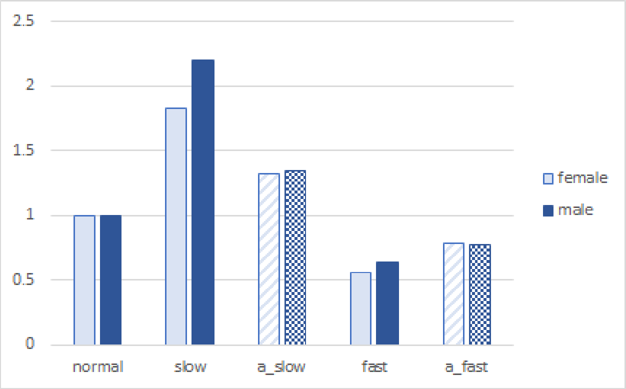


Fig. 4. Average duration ratio of vowels in "normal," "slow," and "fast" for female and male voices. The artificial data is marked "a_slow" and "a_fast."

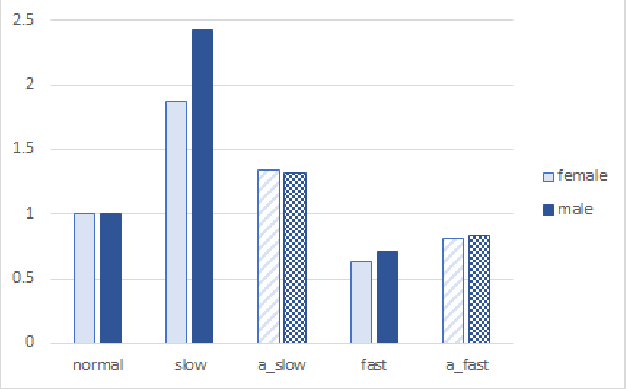


Fig. 5. Average duration ratio of consonants in "normal," "slow," and "fast" for female and male voices. The artificial data is marked "a_slow" and "a_fast."

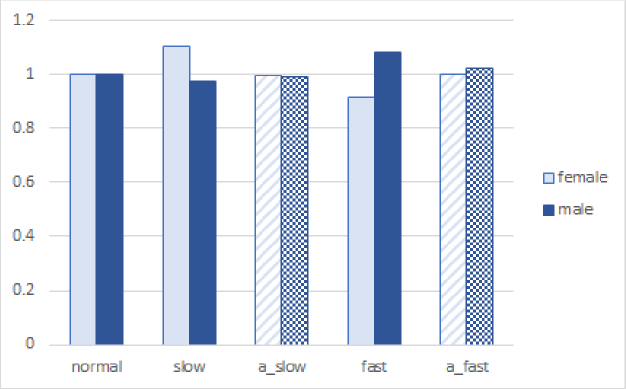


Fig. 6. Average power ratio of all phonemes in "normal," "slow," and "fast" for female and male voices. The artificial data is marked "a_slow" and "a_fast."

condition. These phenomena were identified in both female and male speakers.

We further investigated the differences between the average duration of the vowels or consonants. Here we expected no difference in the duration of the consonants between the natural and artificial data, but there was a large difference in the vowels. Fig. 4 shows the average duration ratio of the vowels in female and male voice, and Fig. 5 shows the average duration ratio of the consonants in female and male voice. However, from these figures, we identified a difference in the vowels and the consonants in female voice. A similar tendency is also indicated in the male voices for their average duration ratio of the vowels and consonants.

Last, we investigated the average power of all the phonemes between the natural and artificial data. Fig. 6 shows the results for both female and male voices. We found no significant difference in the change in the power at different speaking rates. In other words, in reading-style speech, a change in the speaking rate does not change the volume.

In summary, since the difference in vowels and consonants tends to have the same phenomena, distinguishing between vowels and consonants may be unnecessary when controlling the speaking rate in English speech. We also confirmed that we do not have to address the power for controlling the speaking rates since almost no change in it occurs due to the change in the speed.

V. Experiments

Next we discuss the evaluation of our proposed model. Given an input sentence, we first generated full context labels using part of the tools from the HMM/DNN-based Speech Synthesis System (HTS) [15], [16]. After that, we included speaker variation information into the labels based on the two proposed methods described in Section 3. In proposed method-1, we added a discrete label ("N,""S," and "F") to the phoneme level, but in proposed method-2, we added a continues value in the utterance level and trained the GAN-TTS based on these data. For comparison, we applied sox that changed the speed on the synthesized output produced with the "normal" data.

Here we used a preference (AB) test to evaluate the performance and subjectively assessed the speech's naturalness. 11 subjects (7 males, 4 females), who have TOEIC[1] score higher than 700 points (daily conversational level), participated in the experiments. We randomly gave them two speech utterances and asked them to answer which voice sounded more natural: voice A, voice B, or no difference. Fig. 7 compares the proposed method-1 and the sox baseline, and Fig. 8 shows the contrast between proposed method-2 and the sox baseline. Proposed methods-1 and -2 are more natural than the baseline. Fig. 9 compares proposed methods-1 and -2, where there is no significant difference in the quality of the synthesized speech among the proposed methods.

Next we evaluated the effectiveness of the phoneme-level speaking rate variation. As we discussed in the introduction, we generally talk much more quickly during an emergency and generally slow down for emphasis. We generated the following three types of speech: (1) normal, (2) constantly faster, and (3) slowing down on specific words or phrases for emphasis. However, if the speech speed is correctly controlled

[1] Test of English for International Communication (TOEIC) – https://www.ets.org/toeic/

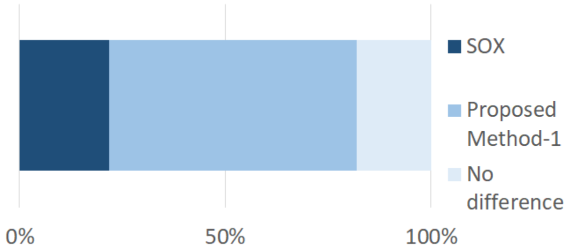


Fig. 7. ABX preference test on naturalness: The baseline versus the proposed method-1.

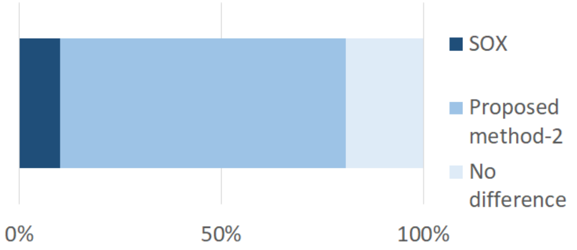


Fig. 8. ABX preference test on naturalness: The baseline versus the proposed method-2.

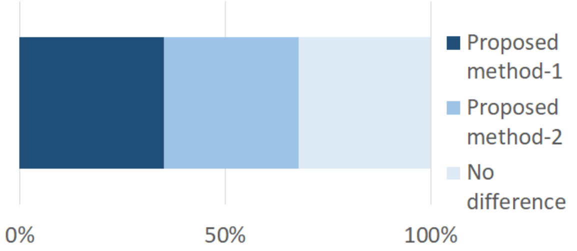


Fig. 9. ABX preference test on naturalness: The proposed method-1 versus the proposed method-2.

by GAN-TTS, then only the relevant part sounds as if it were emphasized. If partial speech speed control is impossible, the emphasis part cannot be identified. Since we found no difference in the quality between the two proposed methods in how the speech speed information was given, we performed this experiment with the phoneme extension method or our proposed method-1.

Again, a preference (AB) test was used to evaluate the performance. We presented the subject a text with emphases marked (in bold) on certain words, and two speech utterances that presented randomly. Then, we asked the subjects to answer which voice that sounds more emphasized. Figure 10 shows the results of the proposed method in comparison with consistently fast speech. The results reveal that the proposed method can change the speaking rate appropriately. In other words, it was possible to make an utterance that emphasized only on a specific word or phrase in the sentence. However, according to the Figure 11, there is no difference between the proposed method and the “Normal” speech. This was because the recorded ‘Normal” speech was slower than standard natural reading speech, so it sounds like the whole sentence was emphasized, and therefore there was no difference with the proposed method.

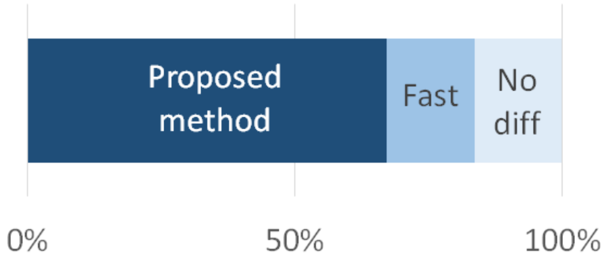


Fig. 10. ABX preference test on speaking rate: Baseline that consistently generates fast speech rate versus the proposed method that changes the speaking rate on emphases word.

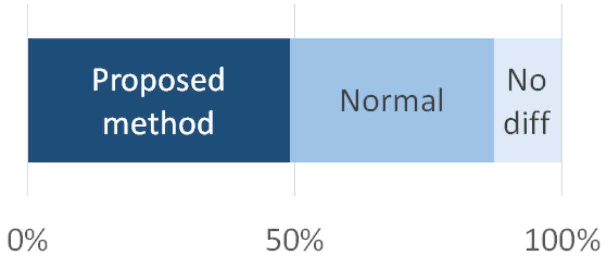


Fig. 11. ABX preference test on speaking rate: Baseline that consistently generates normal speech rate versus the proposed method that changes the speaking rate on emphases word.

Again, we used a preference (AB) test to evaluate the performance. We presented the subjects with a text whose emphasis was marked (in bold) on certain words and randomly presented in two speech utterances. Then the subjects answered which voice received more emphasis. Figure 10 compares the results of our proposed method with consistently fast speech. The former appropriately changed the speaking rate. In other words, we can create an utterance that only emphasizes a specific word or phrase in a sentence. However, based on Fig. 11, perhaps no difference can be found between the proposed method and “normal” speech. Since the recorded “normal” speech was slower than the standard natural reading speech, it sounds like every sentence was emphasized, and therefore no difference was identified with the proposed method.

VI. Conclusion

We proposed a GAN-TTS that controls the variation of the speaking rate at the phoneme level to allow it to change within utterances. We proposed two methods and experimentally verified their usefulness. Our proposed method, which is more natural than artificially manipulating the waveform of synthetic speech, can appropriately perform speaking rate variation at the phoneme level. To some degree, emphasis can be shown by slowing down the speaking rate of certain words. In the future, we will further investigate the possibility of varying the speaking rate to entrain human speech as a dialog partner.

ACKNOWLEDGMENT

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

REFERENCES

[1] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "Atr υ-talk speech," in *Proceedings of ICSLP*, 1992, pp. 483–486.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, 1996, pp. 373–376.

[3] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proceedings of ICASSP*, 1995, pp. 660—-663.

[4] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[6] A. 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.

[7] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 84—-96, 2018.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of Eurospeech*, 1999, pp. 2347—-2350.

[9] J. Yamagishi, T. Masuko, and T. Kobayashi, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of Special Workshop in Maui*, 2004.

[10] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.

[11] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[12] J. Park, K. Han, Y. Jeong, and S. W. Lee, "Phonemic-level duration control using attention alignment for natural speech synthesis," in *Proceedings of ICASSP*, 2019, pp. 5896–5900.

[13] "SoX: Sound eXchange, the Swiss Army knife of audio manipulation," http://sox.sourceforge.net/.

[14] J. Kominek and A. Black, "The CMU arctic speech databases," in *Proceedings of SSW*, 2004.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2000 IEEE International Conference on*. IEEE, 2000, p. 1315–1318.

[16] "The HMM-based speech synthesis system (HTS)," http://hts.ics.nitech.ac.jp.