Speaking Rate Control by HiFi-GAN using Feature Interpolation

Detai Xin[1,a)] Shinnosuke Takamichi[1,b)] Takuma Okamoto[2,c)] Hisashi Kawai[2,d)]
Hiroshi Saruwatari[1,e)]

Abstract: We investigate the possibility of controlling speaking rate by using the HiFi-GAN neural vocoder. Although traditional time-scale modification (TSM) algorithms have been widely applied in real-world applications, their performance and efficiency are relatively low. Recent work of neural vocoder has shown the possibility of synthesizing speech with high fidelity and efficiency. The proposed method inserts an interpolation layer into the HiFi-GAN to control the speaking rate. A signal resampling method and an image scaling method are implemented in the proposed method to warp the mel-spectrogram or hidden features of the neural vocoder. We also design a Japanese speech corpus to evaluate the proposed speaking rate control method. Experimental results of comprehensive objective and subjective evaluations demonstrate that the proposed method can control speaking rate with higher quality and efficiency than a baseline TSM algorithm. We open-source the corpus and give future directions of speaking rate control by a neural vocoder.

Keywords: speaking rate, speaking rate control, HiFi-GAN, neural vocoder, WSOLA

1. Introduction

Speaking rate control is a traditional speech processing task of changing the speaking rate of a given speech. This technology has been widely applied in many real-world applications. Since speaking rate changes speech intelligibility and accessibility, users may need different speaking rates depending on the situation. For instance, a high speaking rate may be needed by visually-impaired people [1] and students who take online courses [2] since they are used to listening to speech with a higher speed. Instead, people who are in an environment with high reverberation [9] or people who have aphasia [4] and elderly adults [12] may need a low speaking rate to understand the content of speech.

The most prevailing method for speaking rate control is time-scale modification (TSM) algorithms [3], [8] based on waveform warping. Basically, such an algorithm first decomposes a waveform into several frames, then warps each frame to the target rate and finally reconstructs the waveform. However, although TSM algorithms have acceptable quality and efficiency, their performance is still not satisfactory. Moreover, recent advances in applying deep neural networks (DNNs) for speech synthesis like neural vocoder [5], [6] demonstrate superior performance on synthesizing speech with high fidelity and efficiency, but since TSM algorithms modify time-domain waveforms directly, it is difficult to combine a DNN-based speech synthesis model with a TSM algorithm. Therefore, a powerful and efficient speaking rate control method that can be seamlessly implemented in DNN-based speech synthesis models becomes necessary. A DNN-based speaking rate control method with multi-speaker WaveNet vocoder had been initially provided and it outperformed conventional TSM-based method and source-filter vocoder [10]. However, the inference speed of the method was quite slow due to the auto-regressive structure and the huge size of the WaveNet model [11]. Most recently, Morrison et al. [7] proposed a speaking rate control method using LPCNet for real-time inference [14]. However, the method was mainly designed for pitch shifting, and the neural vocoder used in their work is different

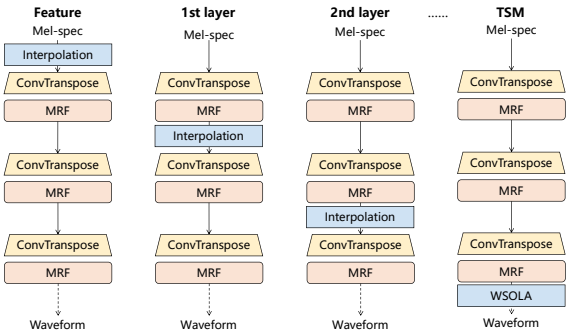


Fig. 1 Diagram of the idea of this work. An interpolation layer is inserted into HiFi-GAN to warp the length of mel-spectrograms or hidden features and control the speaking rate. When interpolating the output waveform, the proposed method becomes the traditional TSM method, which is used as the baseline method. In this work we use WSOLA as the implementation of the baseline method.

[1] The University of Tokyo, Hongo, Tokyo 113-8656, Japan
[2] National Institute of Information and Communications Technology, Kyoto, 619-0289, Japan
[a)] detai_xin@ipc.i.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[c)] okamoto@nict.go.jp
[d)] hisashi.kawai@nict.go.jp
[e)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

Table 1 Mel-cepstral distortion of the converted speech. HiFi-GAN represents synthesized speech without rate changing.

From	To	HiFi-GAN	mel (l)	mel (k)	WSOLA	L1 (l)	L2 (l)	L3 (l)	L4 (l)	L1 (k)	L2 (k)	L3 (k)	L4 (k)
Slow	Norm	1.15	**2.20**	2.22	2.37	3.38	2.74	3.12	3.39	3.35	2.72	3.14	3.40
Slow	Fast	1.15	2.99	**2.96**	3.18	3.85	4.07	4.38	4.92	3.81	4.06	4.40	4.93
Norm	Fast	1.16	**2.74**	2.76	2.91	3.80	3.54	3.88	4.13	3.83	3.53	3.89	4.15
Norm	Slow	1.16	**2.24**	**2.24**	2.37	3.42	2.85	3.05	3.27	3.46	2.85	3.06	3.31
Fast	Slow	1.41	**2.83**	2.87	3.03	3.49	3.68	3.75	3.94	3.48	3.67	3.76	3.98
Fast	Norm	1.41	**2.62**	2.65	2.83	3.62	3.32	3.55	3.78	3.65	3.33	3.56	3.82

from the one used in this work.

In this paper, we investigate the possibility of using a neural vocoder to control speaking rate. We use a recent GAN-based vocoder HiFi-GAN [5] as the vocoder, which can synthesize high-fidelity speech with 0.01 of the real time factor. Our idea is illustrated in **Fig. 1**. The proposed method interpolates mel-spectrograms or hidden features in the possible inner layers of HiFi-GAN to control the speaking rate as [10]. When the interpolation layer is inserted before the generated waveform, the proposed method degrades to a TSM algorithm, which is used as the baseline method of this work. We consider both a bandlimited signal resampling and an image scaling interpolation methods in the experiments. To evaluate the proposed method, we built a corpus with three kinds of speaking rates of unique speakers and conducted experiments on it. Results of comprehensive objective and subjective evaluations demonstrated that the proposed method can synthesize speech with higher fidelity and efficiency than the baseline TSM algorithm. The key contributions of this work are as follows:

- We propose a speaking rate control method by a neural vocoder with high fidelity and efficiency.
- We design a corpus for speaking rate control and give insights through experiments upon this corpus. The corpus is open-sourced in `https://ast-astrec.nict.go.jp/release/speedspeech_ja_2022/download.html`.

2. Proposed Method

2.1 Overview

We use HiFi-GAN [5] as the neural vocoder used in experiments. HiFi-GAN utilizes adversarial training to discriminate synthetic and genuine speech in parallel, thus can synthesize speech with high fidelity and efficiency. As Fig. 1 illustrates, the basic idea of the proposed method is using an interpolation layer to change data length so that the speaking rate of the synthesized speech will be different from the original mel-spectrogram. Since the generator of HiFi-GAN has several similar blocks to upsample the data, it is natural and intuitive to insert an interpolation layer between these blocks.

As mentioned in the previous section, when the interpolation layer is inserted after the last upsampling layer, the proposed method becomes a traditional TSM algorithm. We use the waveform similarity overlapping-add (WSOLA) algorithm [15] as the baseline method. WSOLA can change the speaking rate of the speech while maintaining the periodic patterns of the signal by finding a frame that has the maximal similarity to the current frame. Benefiting from this property, WSOLA is suitable for modifying the time-scale of human speech.

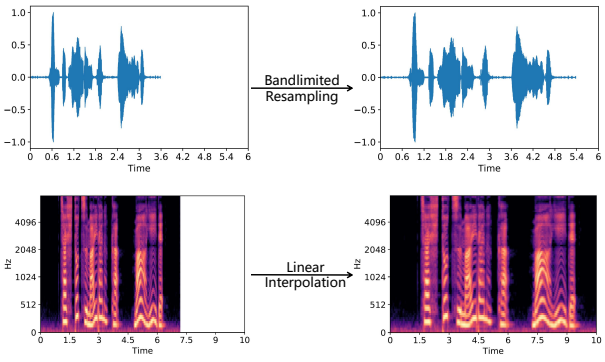


Fig. 2 Examples of bandlimited resampling and linear interpolation. Bandlimited resampling (top-half) is usually used to interpolate one-dimensional waveforms, while linear interpolation (bottom-half) is designed to interpolate two-dimensional data like mel-spectrograms.

2.2 Interpolation Methods

The interpolation layer should change the feature length while maintaining the semantic information within it. In this work we consider two interpolation methods: bandlimited resampling based on kaiser window and linear interpolation for image scaling. Examples of these two methods are illustrated in **Fig. 2**. Bandlimited interpolation is a commonly used resampling method in digital signal processing. Since the outputs of the hidden layers of HiFi-GAN can be regarded as hidden representations of waveforms, we consider that it is reasonable to use this method to interpolate them. Besides, it is also intuitive to interpret hidden features including mel-spectrograms as images that contain two axes time and frequency. Therefore we also use a geometric interpolation method linear interpolation. In the experiments we compare these two methods and show that linear interpolation is better than bandlimited resampling.

3. Experiments

3.1 Setup

To evaluate the proposed method, we designed a Japanese dataset with three kinds of speaking rates. The dataset contains a male and a female speakers. Each speaker speaks 325 utterances at three different rates: slow, normal, and fast, so each speaker has 975 utterances. We name this dataset NICT-SpeedSpeech. We randomly picked up 45 utterances for each speaker in which each rate has 15 utterances. The texts of the 15 utterances of each speaker between different rates were kept to be the same so that we could measure the mel-cepstral distortion in the later experiments. For all speech of each speaking rate, we converted them to the rest two rates and regarded the speech of the target rate as the ground truth data. Since every utterance has a different speaking rate, we computed the conversion fac-

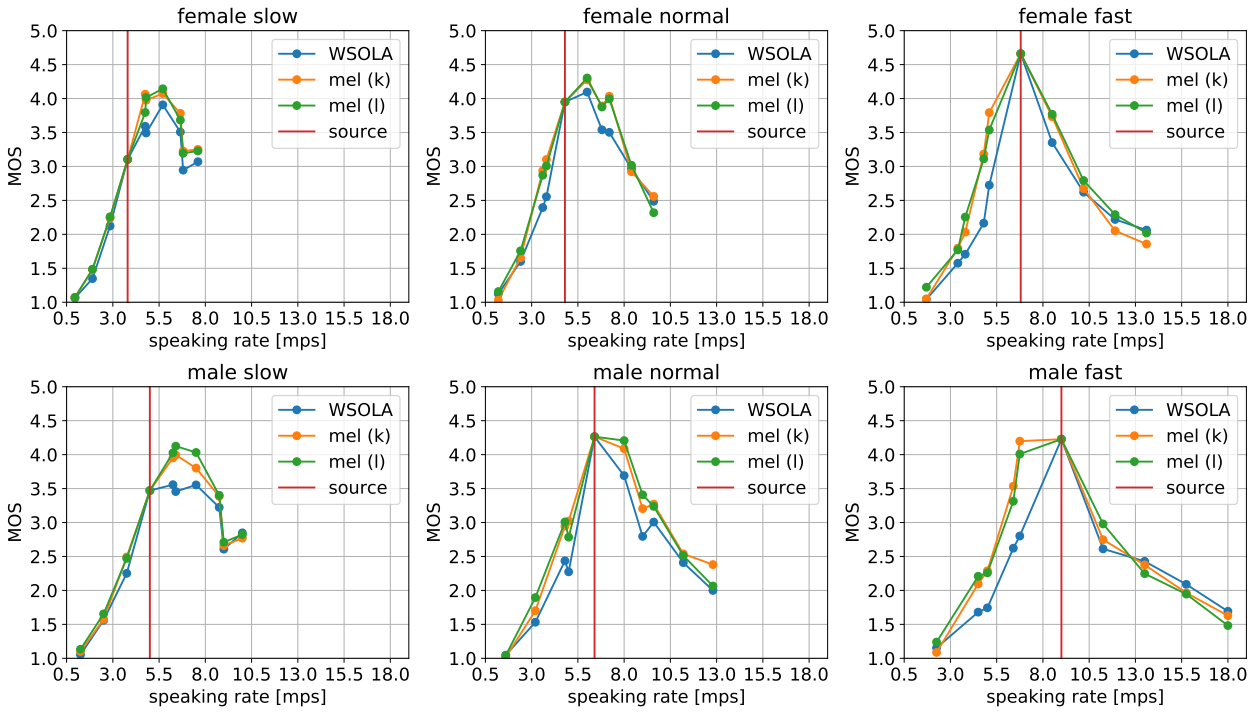


Fig. 3 Results of naturalness MOS evaluation. The title of each sub-figure represents the speaker and the speaking rate of the source utterances. The read vertical line (original) represents the exact speaking rate of the source utterances.

tor for each utterance separately. Denoting the durations of the source and the target utterances with the same text as t_{src} and t_{tgt}, the conversion factor f is defined as $t_{\text{src}}/t_{\text{tgt}}$. In the subjective evaluations, we also converted the speech with standard factors $\{0.25, 0.5, 0.75, 1.25, 1.5, 1.75, 2.0\}$ that are widely used in real-world applications.

We used a pretrained universal HiFi-GAN model[*1] to convert mel-spectrograms into time-domain waveforms. This model was trained on multilingual datasets so it can synthesize Japanese without further training. Since the generator of the pretrained HiFi-GAN has four upsampling layers, it is possible to insert the interpolation layer into 5 places (mel-spectrogram and layer $1, 2, 3, 4$). All combinations of interpolation methods and places were tried. As a result, there were 11 methods including 10 proposed methods and 1 baseline method in the experiments. For simplicity, we denote these methods as mel (l/k), L{1,2,3,4} (l/k) and WSOLA, where l denote linear interpolation and k denote bandlimited resampling based on kaiser window. We used the bandlimited resampling class in torchaudio[*2] and the linear interpolation function of pytorch[*3] as the implementations.

For the WSOLA algorithm we used audiotsm[*4] package as the implementation.

[*1] https://github.com/jik876/hifi-gan
[*2] https://pytorch.org/audio/stable/transforms.html#torchaudio.transforms.Resample
[*3] https://pytorch.org/docs/stable/generated/torch.nn.functional.interpolate.html
[*4] https://github.com/Muges/audiotsm

Table 2 Speaking rate (mora per second) of the proposed NICT-SpeedSpeech dataset

Speaker	Fast	Normal	Slow
Male	9.0 ± 0.3	6.4 ± 0.4	5.0 ± 0.2
Female	6.8 ± 0.6	4.8 ± 0.3	3.8 ± 0.3
Average	7.9 ± 1.2	5.6 ± 0.9	4.4 ± 0.7

3.2 Objective Evaluations

3.2.1 Speaking Rate of Ground Truth Data

We first measured the exact speaking rate of the dataset by computing mora per second of the speech. This was computed by dividing the mora number by the voiced length of each utterance. The voiced frames were detected by using voice activity detection[*5]. The result is shown in **Table 2**. It can be seen that the speaking rate of the fast speech is almost 2-time faster than the one of the slow speech. Also, the male speaker speaks faster than the female speaker.

3.2.2 Mel-cepstral Distortions

We then measured mel-cepstral distortions (MCDs) between the converted and ground truth speech to evaluate each method. Utterance pairs were aligned by dynamic time warping. The result is shown in **Table 1**. The mel-spectrogram linear interpolation method (mel (l)) obtained the best performance. Besides, it can be observed that interpolating mel-spectrograms is better than interpolating hidden features of inner layers. The baseline method (WSOLA) also obtained relatively good performance but is still worse than the two mel-spectrogram interpolation methods. Surprisingly, all the methods of interpolating hidden features of inner layers have poor performance. After a preliminary analysis, we found that the fundamental frequency (F0) of the ut-

[*5] https://github.com/wiseman/py-webrtcvad

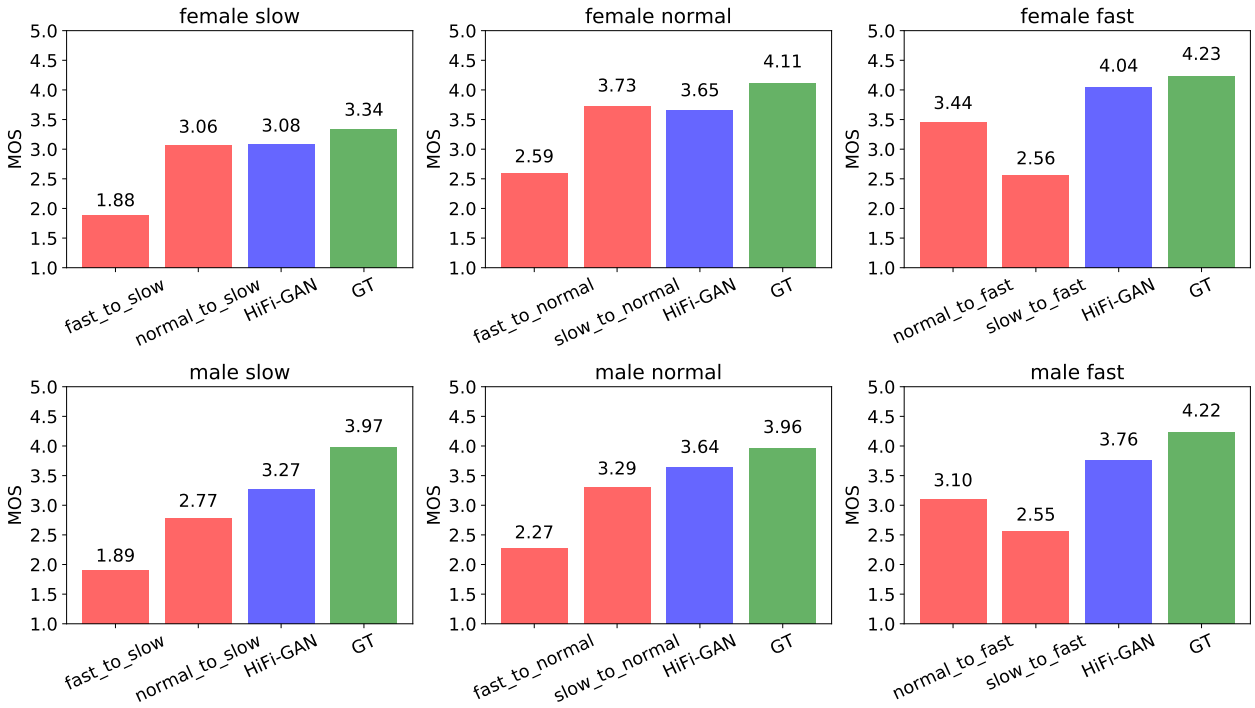
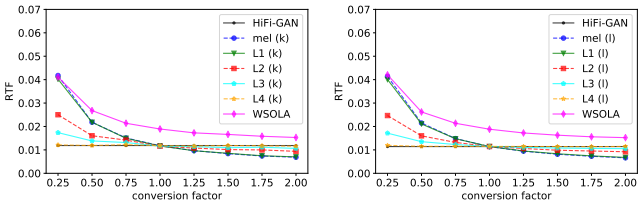


Fig. 4 Results of naturalness MOS evaluation with fixed target speaking rate. The title of each sub-figure represents the speaker and the target speaking rate. The red bar denotes the mel (l) method which obtained the best performance in the previous evaluations.



(a) Bandlimited resampling (b) Linear interpolation

Fig. 5 Results of efficiency evaluation. The efficiency measured by RTF is compared for (a) proposed method using bandlimited resampling and WSOLA, (b) proposed method using linear interpolation and WSOLA. HiFi-GAN represents the efficiency of the model itself without changing the speaking rate.

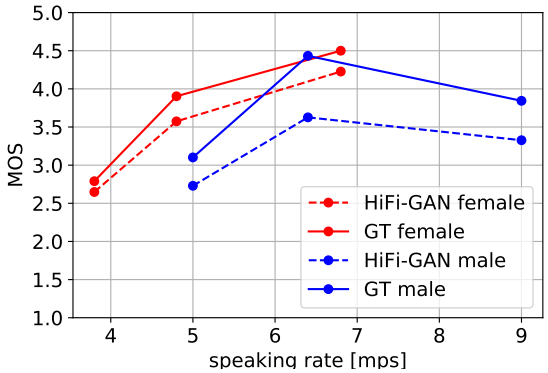


Fig. 6 Results of naturalness MOS evaluation for natural speech (GT) and synthetic speech (HiFi-GAN) of each speaker under different speaking rate without speaking rate control.

terances converted by these methods were destructed, though the semantic information was well preserved. We assume this is because the CNN architecture of HiFi-GAN makes it depend on the hidden feature length to work, thus changing the feature size will influence the behavior without further training.

3.2.3 Real Time Factors

To evaluate the efficiency of each method, we then computed real time factors (RTFs) for each method using an NVIDIA GeForce RTX 2080 Ti GPU card. The computation time is defined as the summation of the generation time of HiFi-GAN and the speaking rate conversion time of each method. For simplicity in this evaluation we only used standard factors to convert all slow, normal and fast utterances but did not convert between them. The result is shown in **Fig. 5**. It can be seen that all proposed methods have better efficiency than the baseline WSOLA method. The mel-spectrogram interpolation methods (mel (l/k)) which obtained the best performance in the MCDs evaluation have the best performance when the conversion factor is greater than 1 but obtained the worst performance among the proposed methods when the conversion factor is less than 1. This is because the four upsampling blocks in HiFi-GAN will magnify the data length changing effect, therefore the shortened or elongated data will become shorter or longer after being processed by the blocks, and further increase or reduce the computation time.

3.3 Subjective Evaluations

3.3.1 Speech Naturalness of HiFi-GAN

We first conducted a standard 5-scale mean opinion score (MOS) test to evaluate the synthetic speech generated by HiFi-GAN without rate control. Totally 100 listeners joined in the test. Each utterance was evaluated by 10 listeners. The result is shown in **Fig. 6**. It can be seen that the listeners tend to rate low scores for speech with uncommon speaking rates, even for the natural

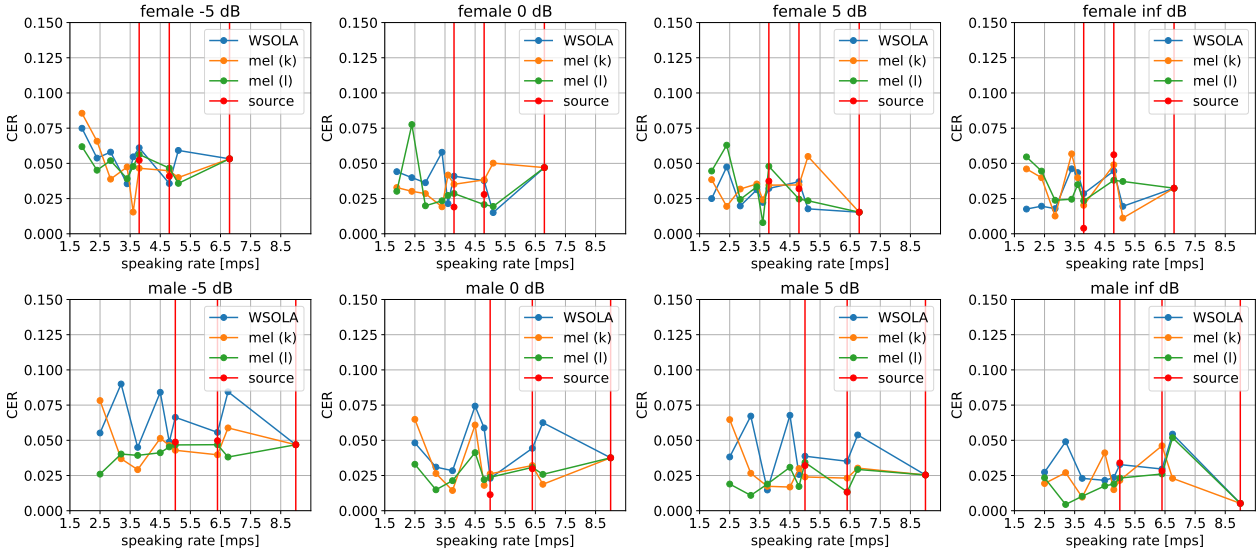


Fig. 7 Results of intelligibility evaluation. The title of each sub-figure represents the speaker and the SNR. The red vertical line represents speech without changing the speaking rate.

speech. Also, there is still a quality gap between the natural and synthetic speech, which we assume is because we didn't fine-tune the model on the dataset.

3.3.2 Speech Naturalness of Converted Speech

Next, we evaluated the naturalness of the converted speech by a MOS test. On the basis of the results of the objective evaluations, we only selected three methods: WSOLA, mel-spec interpolation using linear interpolation or bandlimited resampling in this test.

In the first evaluation, we intended to see the influences of the source speaking rates on the MOS. We mixed all utterances regardless of their original speaking rates. Totally 1500 listeners joined in the test. Each utterance was evaluated by 30 listeners. The result is illustrated in **Fig. 3**, each column represents a source speaking rate. It can be observed that if the speaking rate is in a common range [4, 10], the two mel-spectrogram interpolation methods are better than WSOLA and the linear interpolation method is slightly better than the kaiser window method, which is consistent with the conclusions we got from the objective evaluations. This implies that it is easy to slow down the utterances with high speaking rates or speed up the utterances with low speaking rates. Second, when the speaking rates become extremely slow or fast, the MOS scores become very low and it is hard to distinguish which method is better for the listeners. We assume that there are two possible reasons for this problem. One possible reason is that the pretrained HiFi-GAN model was trained on a corpus with normal speaking rates, so it cannot synthesize speech with slow or fast speaking rates. Another possible reason is that the listeners are not used to listening to speech with uncommon speaking rates. But since it can be observed that the listeners rate low scores even for natural speech with uncommon rates from the previous section (Section 3.3.1), we conclude that the latter reason is more appropriate to explain the result.

In the second evaluation, we instead aimed to know the performance of the proposed method when the target speaking rate is fixed. Based on the first evaluation, we only used the utterances converted by the mel (l) method. For simplicity we did not use the utterances converted using standard factors. We categorized all utterances into 6 groups by their speaker and target speaking rate, and conducted a MOS test for each group. Each test has 35 listeners. The result is illustrated in **Fig. 4**. It can be seen that when the source and target speaking rates are quite different, e.g. slow to fast, the performance becomes bad. Also, the listeners tend to rate high scores for utterances with normal speaking rates, which is consistent with the conclusion of the previous evaluation. In the task of converting female slow utterances to normal utterances, the performance of the proposed method even surpasses HiFi-GAN, which demonstrates the effectiveness of the proposed method. Although when the target speaking rate is slow or fast, the performance of the proposed method is not as good as HiFi-GAN, in the later section we show that the proposed method can obtain a low character error rate in an intelligibility test, which demonstrates that the listeners can at least understand the converted utterances even if they are not used to listening to such speech.

3.4 Speech Intelligibility

Since speech with extremely fast or slow speaking rates will impede listeners from judging the naturalness, we instead evaluated the speech intelligibility of such speech to see whether the listeners can at least understand the speech or not. We first excluded all utterances whose transcriptions include katakana, then randomly picked up 150 utterances for each speaker with 50 utterances for each speaking rate. In this experiment we only converted the utterances into lower speaking rates, e.g. fast to slow. After conversion, we added random noises from the DEMAND dataset [13] into all utterances with {-5, 0, 5, inf} dB SNRs for each speaker, where inf dB represents clean speech. To reduce the difficulty, we did not use those noises with human voices like OMEETING. The noise category was kept to be the same for utterances with the same text of different methods. Totally 2000 listeners joined in the intelligibility test. Each utterance was evaluated by 40 listeners. The listeners were required to transcript

the utterances. After that, the average character error rate (CER) was computed for each method. The result is shown in **Fig. 7**. It can be seen that in most cases the two mel-spectrogram interpolation methods (mel (l/k)) are better than WSOLA. Also, the linear interpolation method is slightly better than the bandlimited resampling method, which is consistent with the conclusions of the previous evaluations. All in all, the proposed feature interpolation methods have better efficiency and quality performance than the baseline TSM algorithm.

4. Conclusions

This paper described a method for speaking rate control by HiFi-GAN using feature interpolation. The idea of the proposed method is inserting a feature interpolation layer into the model to change the data length and control the speaking rate. A Japanese corpus with various speaking rates was designed to evaluate the proposed method. Results of subjective and objective evaluations demonstrated that the proposed method of mel-spectrogram interpolation using linear interpolation had better efficiency and quality performance than the baseline TSM algorithm WSOLA. The future work will be testing the generalization ability of the proposed method on other neural vocoders.

Acknowledgments This work is supported by JST SPRING, Grant Number JPMJSP2108, and is also supported by JSPS KAKENHI 19H01116, 21H04900.

References

[1] Bragg, D., Bennett, C., Reinecke, K. and Ladner, R.: A large inclusive study of human listening rates, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (2018).
[2] Dodson, E. M. and Blinn, C. R.: How Will COVID-19 Change Forestry Education? A Study of US Forest Operations Instructors, *Journal of Forestry* (2021).
[3] Driedger, J. and Müller, M.: A review of time-scale modification of music signals, *Applied Sciences*, Vol. 6, No. 2, p. 57 (2016).
[4] Hux, K., Brown, J. A., Wallace, S., Knollman-Porter, K., Saylor, A. and Lapp, E.: Effect of text-to-speech rate on reading comprehension by adults with aphasia, *American journal of speech-language pathology*, Vol. 29, No. 1, pp. 168–184 (2020).
[5] Kong, J., Kim, J. and Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis, *Advances in Neural Information Processing Systems*, Vol. 33 (2020).
[6] Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y. and Courville, A. C.: MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis, *Advances in Neural Information Processing Systems*, Vol. 32 (2019).
[7] Morrison, M., Jin, Z., Bryan, N. J., Caceres, J.-P. and Pardo, B.: Neural Pitch-Shifting and Time-Stretching with Controllable LPCNet, *arXiv preprint arXiv:2110.02360* (2021).
[8] Müller, M.: *Fundamentals of music processing: Audio, analysis, algorithms, applications*, Springer (2015).
[9] Nábělek, A. K., Letowski, T. R. and Tucker, F. M.: Reverberant overlap-and self-masking in consonant identification, *The Journal of the Acoustical Society of America*, Vol. 86, No. 4, pp. 1259–1265 (1989).
[10] Okamoto, T., Matsubara, K., Toda, T., Shiga, Y. and Kawai, H.: Neural speech rate conversion with multispeaker WaveNet vocoder, *Speech Communication* (2022, accepted, in press).
[11] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: Wavenet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499* (2016).
[12] Rudner, M., Rönnberg, J. and Lunner, T.: Working memory supports listening in noise for persons with hearing impairment, *Journal of the American Academy of Audiology*, Vol. 22, No. 3, pp. 156–167 (2011).
[13] Thiemann, J., Ito, N. and Vincent, E.: DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, *Proc. Meetings Acoust*, pp. 1–6 (2013).
[14] Valin, J.-M. and Skoglund, J.: LPCNet: Improving neural speech synthesis through linear prediction, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5891–5895 (2019).
[15] Verhelst, W. and Roelands, M.: An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech, *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, IEEE, pp. 554–557 (1993).