

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

Mayuko Okamoto¹, Sakriani Sakti^{1,2}, and Satoshi Nakamura^{1,2}

¹ Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan

² RIKEN, Center for Advanced Intelligence Project AIP (RIKEN AIP), Japan



Background

- Text-to-speech Synthesis (TTS)
 - Development of TTS continues to advance
 - Able to produce speech with high degree of intelligibility
 - Naturalness of generated speech has improved

However, for spoken dialogue applications:

- The generated speech is still too monotonous
 - Lacks variety and liveliness found in natural speech
- Humans
 - Vary their speaking rate
 - Tend to slow down to emphasize words
- For more natural spoken dialogue, TTS that can control speaking rate is better

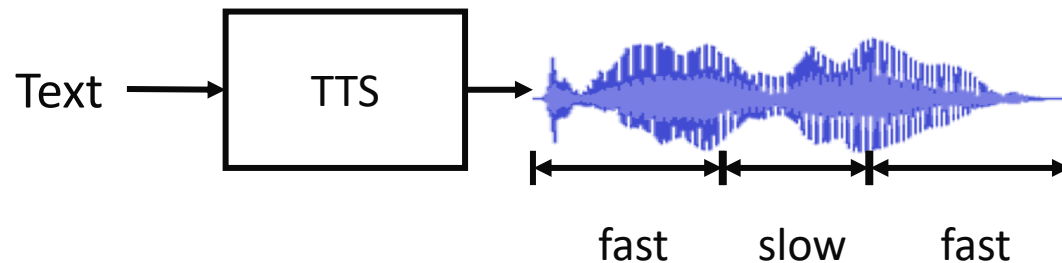


Related Work

- Several studies have addressed speaking style variations
[Yoshimura et al., 1999, Yamagishi et al., 2004]
 - Most existing studies were based on HMM-based TTS
 - Few studies have addressed the speaking rate issue
- Recent studies with seq2seq deep learning
[Wang et al., 2018] Global style tokens for Tacotron
 - Varying speed and speaking style
 - Information is stored globally
 - Controlling at phoneme level is difficult
[Park et al., 2019] Phoneme level duration model in seq2seq
 - Required phoneme input instead of text
 - Only evaluated using simulated data
 - Did not consider the effect on listeners

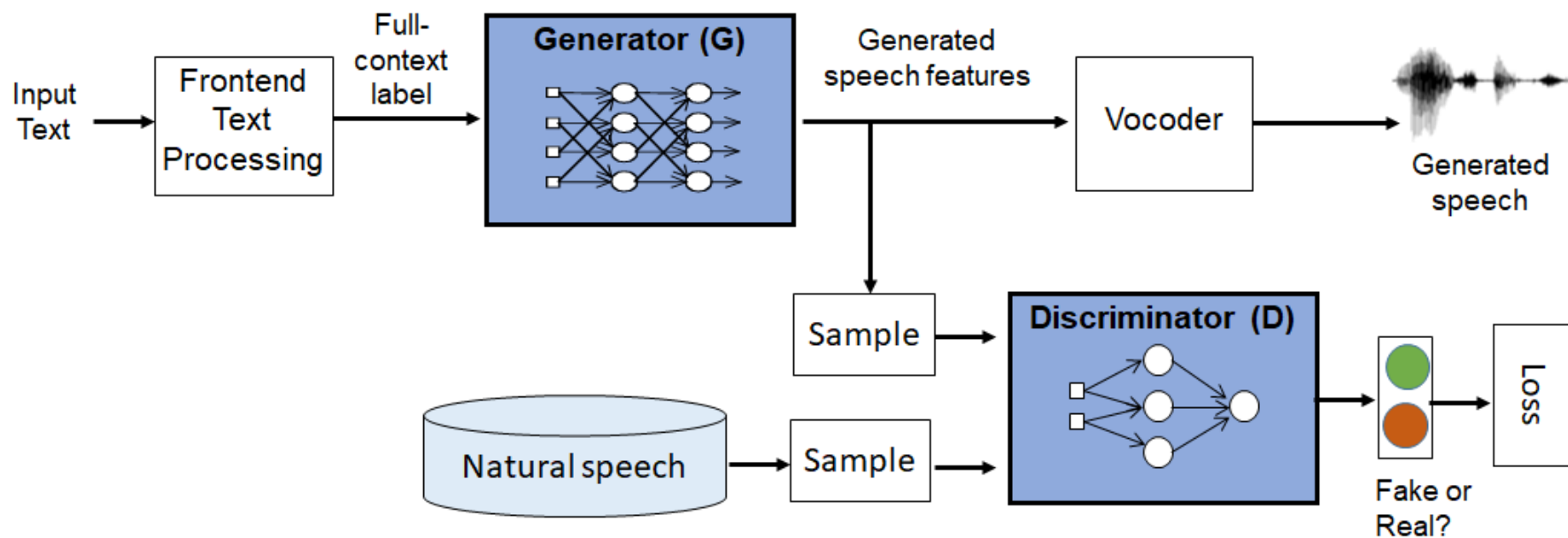
Proposed Approach

Considering speech synthesis that can control utterance features within a conversation

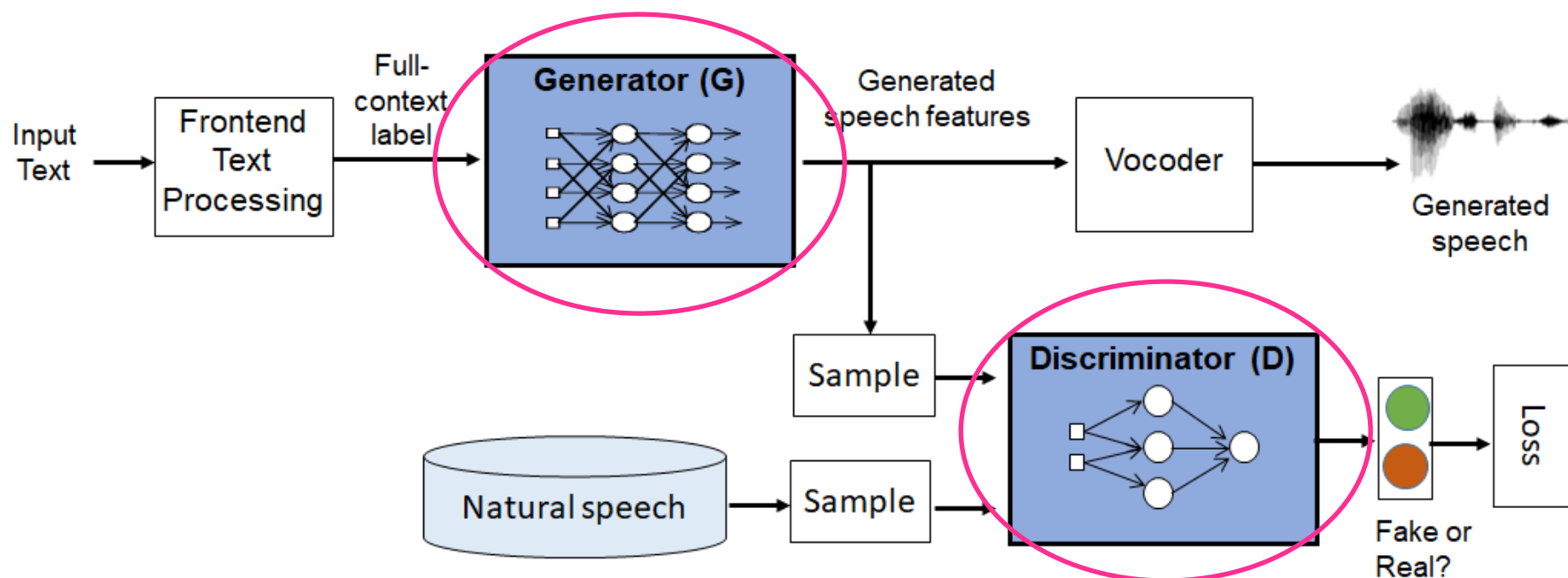


- Construct natural speech corpora
- Analyze differences in speaking rates
- Use TTS based on generative adversarial networks (GAN-TTS) to improve quality of synthesized speech
- Enable GAN-TTS to generate speech waveform with phoneme-level speaking rate variations
- Investigate the effect of speaking rate variation on listeners

Overall Architecture



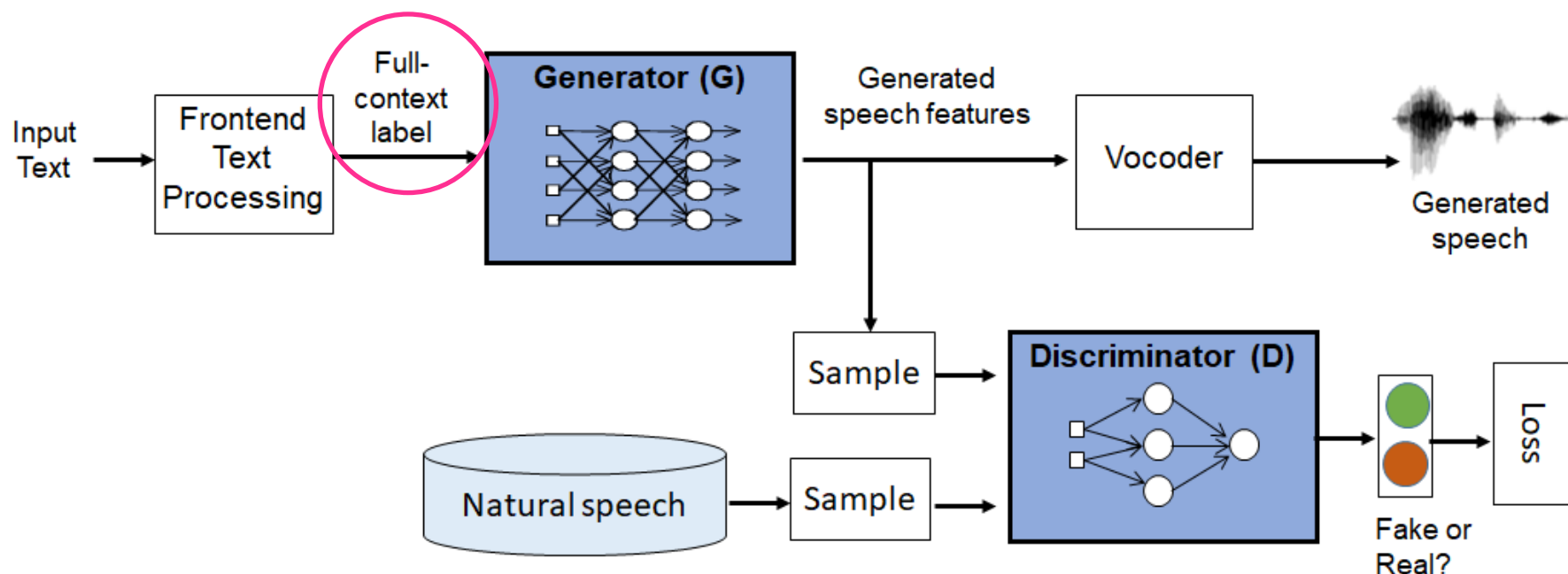
Overall Architecture



GAN consists of 2 networks:

- **Generator (G)**
Learns to create speech output that causes the Discriminator D to misrecognize the generated result as natural speech
- **Discriminator (D)**
Learns to accurately distinguish between natural and synthetic speech produced by the Generator G

Overall Architecture

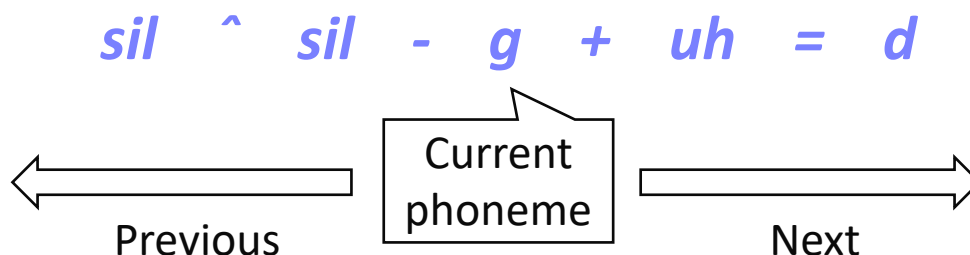


Text	Good morning.
Phoneme	g uh d ...
Triphone	sil-g+uh g-uh+d ...
Full-context label	pentaphone context + syllable context + word level context + phrase level context + utterance level context

Full-context Label

Text	Good morning.
Phoneme	g uh d ...

pentaphone context

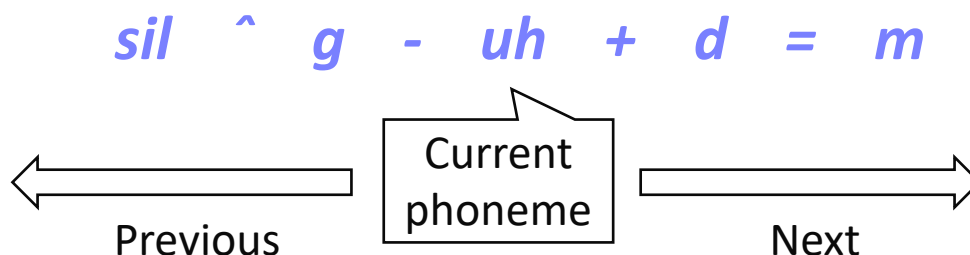


- + **syllable context** (position of “g” in syl, syl stress, #phonemes)
- + **word level context** (position syl in wrd, part of speech, #syllables)
- + **phrase level context** (#syllables and #words in phrase)
- + **utterance level context** (#syllables, #words and #phrase in utterance)

Full-context Label

Text	Good morning.
Phoneme	g uh d ...

pentaphone context



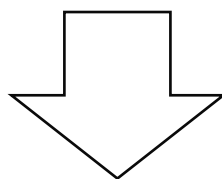
- + **syllable context** (position of “g” in syl, syl stress, #phonemes)
- + **word level context** (position syl in wrd, part of speech, #syllables)
- + **phrase level context** (#syllables and #words in phrase)
- + **utterance level context** (#syllables, #words and #phrase in utterance)

Proposed Method

1. Phonetic symbol expansion

- Add speaking rate information to the phoneme itself

sil^sil-g+uh=d.....



sil^sil-g

<i>S</i>
<i>N</i>
<i>F</i>

+uh

<i>S</i>
<i>N</i>
<i>F</i>

=d

<i>S</i>
<i>N</i>
<i>F</i>

.....

S: Slow, N: Normal, F: Fast

Proposed Method

2. Add ratio of speaking rate

- Specify the speaking rate ratio in the part newly added

[g] sil^sil-g+uh=d... + /K:100

[uh] sil^g-uh+d=m... + /K:100

•
•
•

/K:75 Slow
/K:100 Normal
/K:125 Fast

Data Construction

Based on the CMU ARCTIC database (1132 utterance)

- A) Original utterances
- B) 0.75 x Speaking rate
- C) 1.25 x Speaking rate







→ Sample Data

Recorded natural speech, spoken by one male and one female speaker.

Speakers were asked to record each utterance three times at different speaking rates, in as natural a manner as possible.

→ normal, slow, fast

Data Sample

	female	male
slow		
normal		
fast		

We analyzed our dataset regarding three characteristics:

- Utterance length
- Vowel and consonant length
- Power

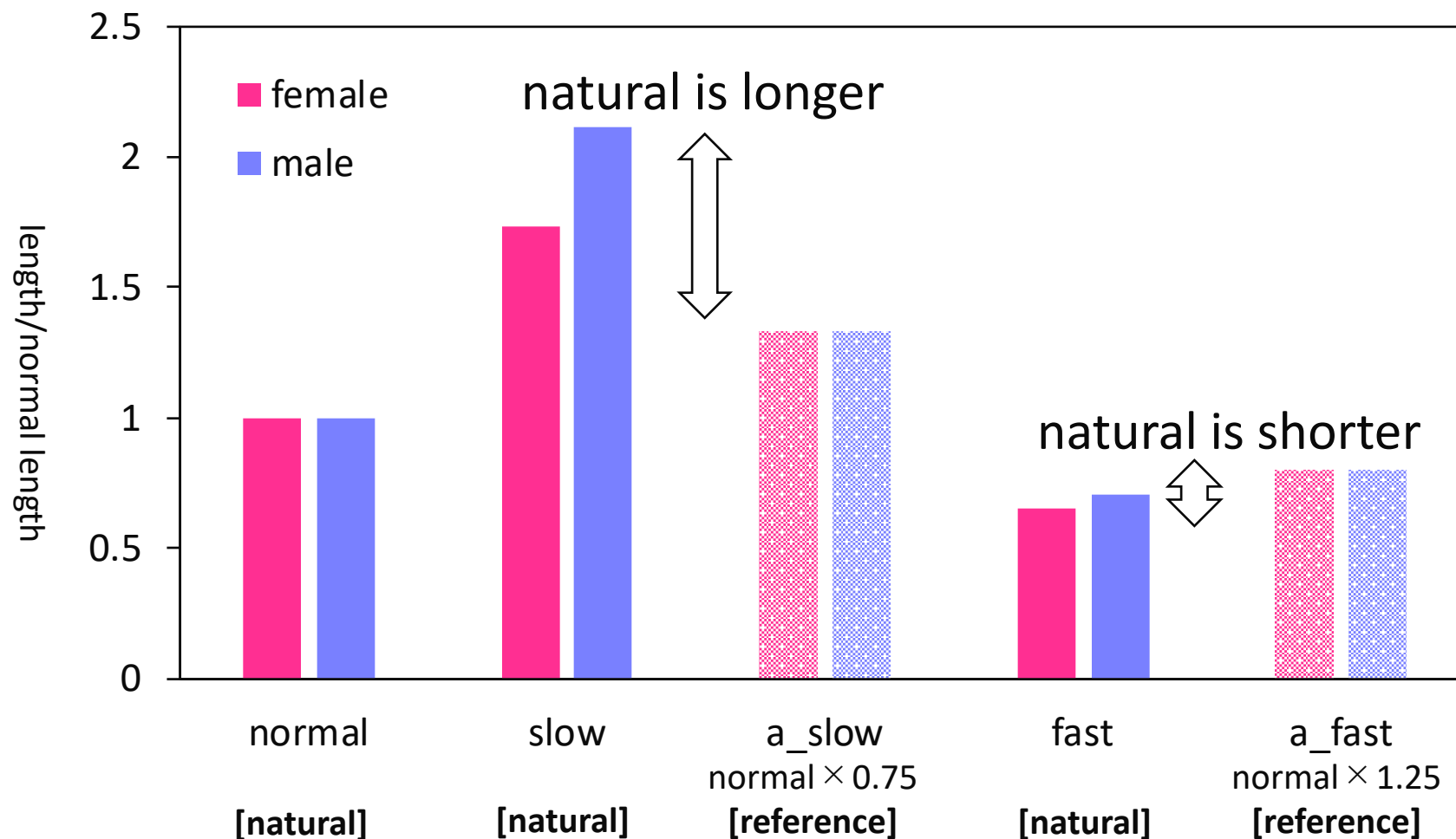
All results are displayed as a ratio with Normal as 1

Reference data → light color

- a_slow = normal speaking rate \times 0.75
- a_fast = normal speaking rate \times 1.25

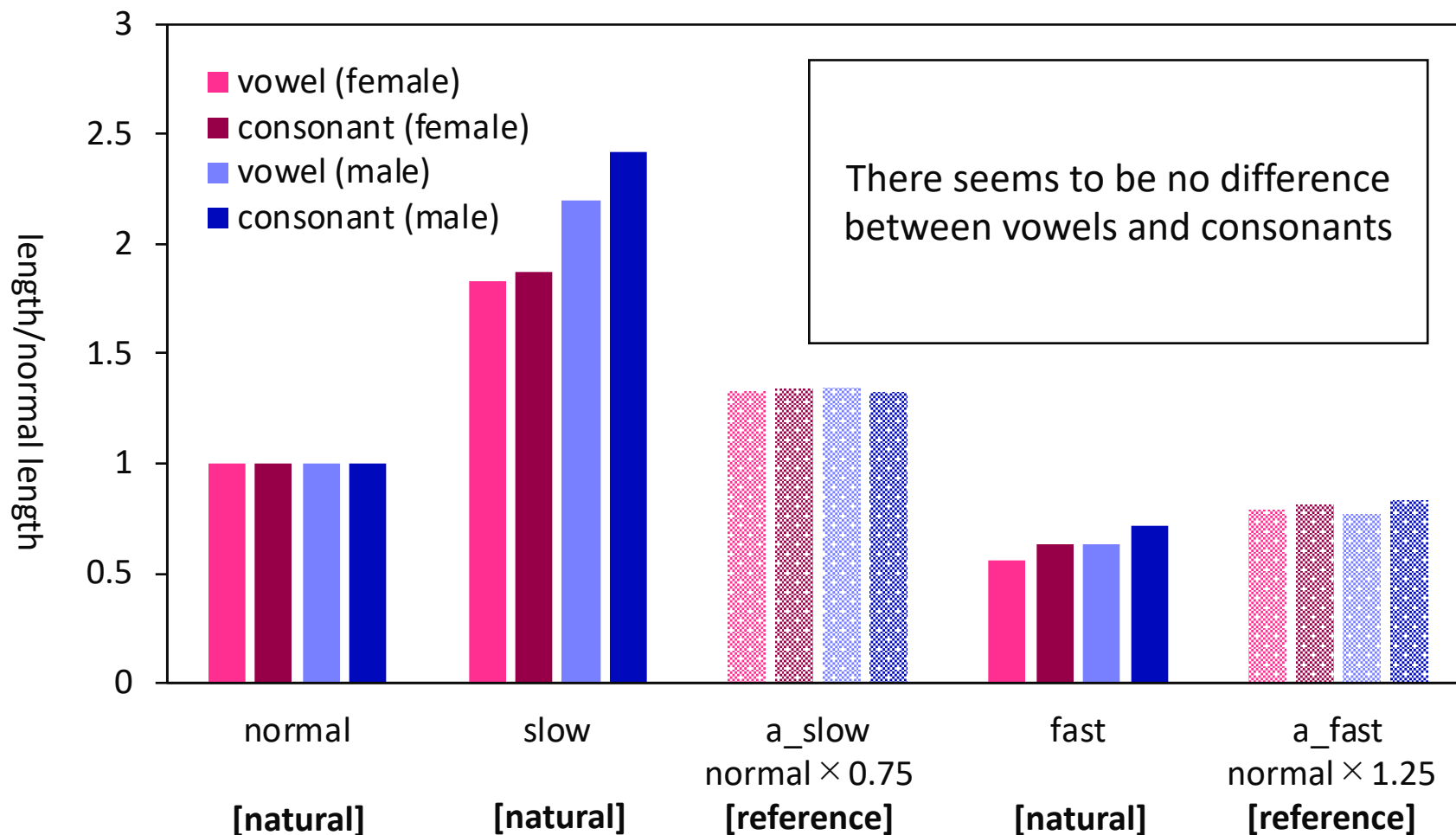
Analysis: Utterance length

Comparing average utterance length
of natural and reference speech



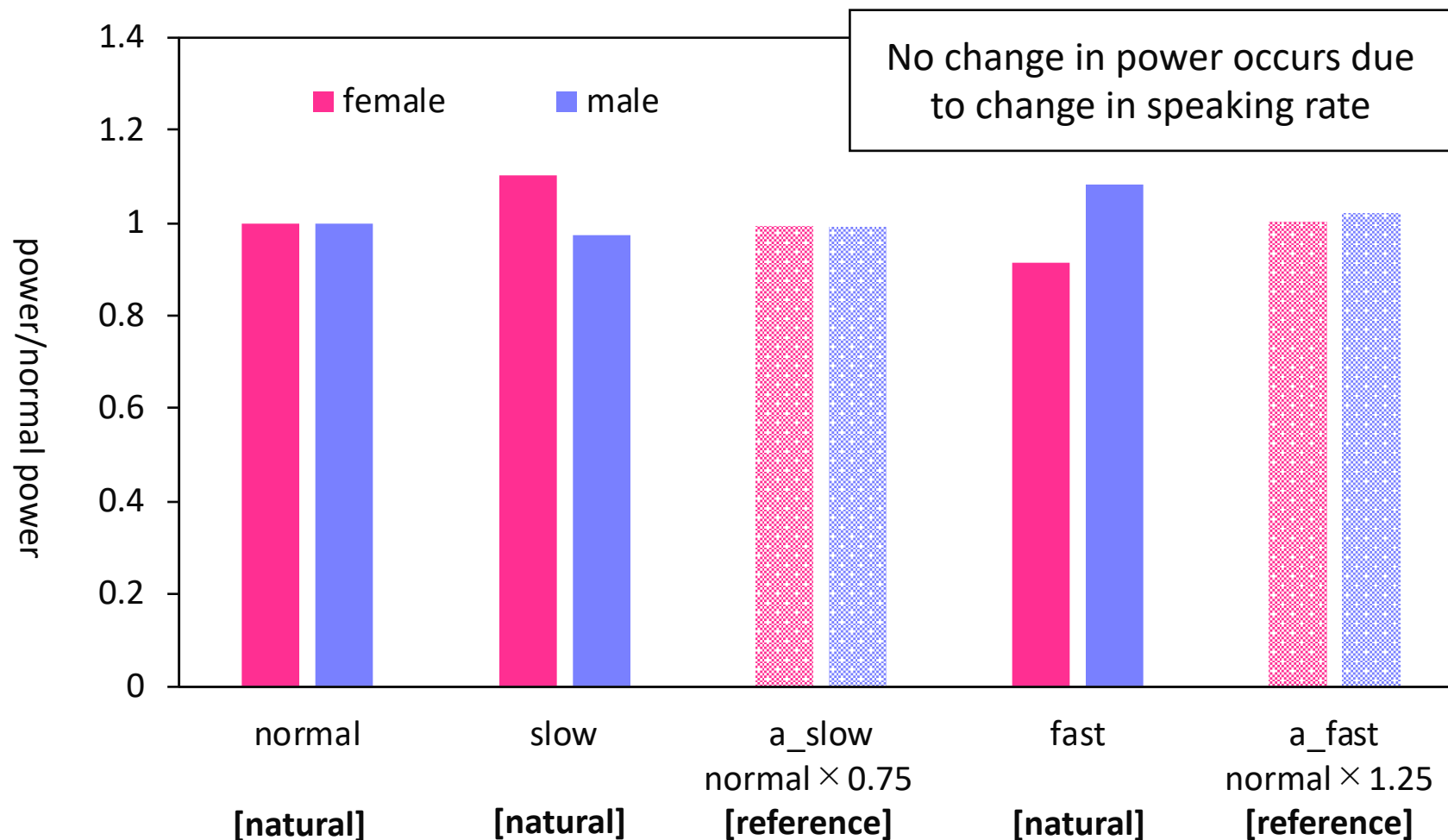
Analysis: Vowel and consonant length

Comparing average vowel and consonant length when the speaking rate changed



Analysis: Power

Comparing average power when the speaking rate changed



Experimental Set-up

We conducted two experiments to confirm the usefulness of the proposed method

- Subjective assessment of the naturalness of generated speech
 - Does speech generated using the proposed method sound more natural than post-processed speech?
- Effectiveness of the phoneme level speaking rate variation
 - Does the proposed method change the speaking rate at the phoneme level?

Subjective preference test









- 11 subjects (7 men, 4 women)
- TOEIC score 700 or higher

Experiments 1: Methods

- **Subjective assessment of the naturalness of generated speech**

- Does speech generated using the proposed method sound more natural than post-processed speech?

- **Generated 3 types of speech**

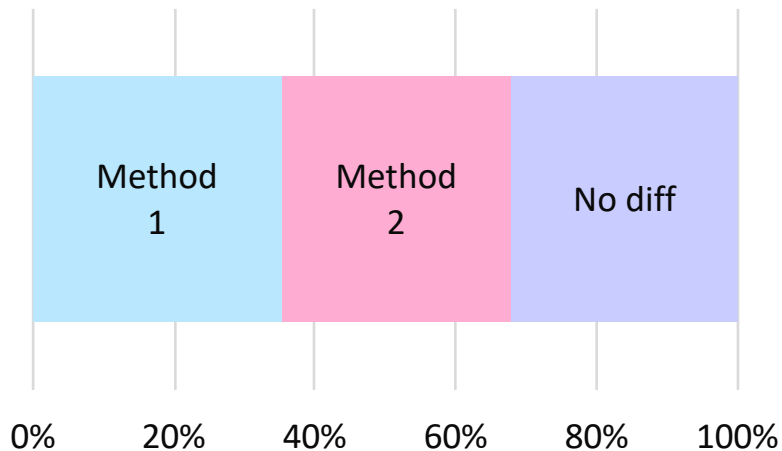
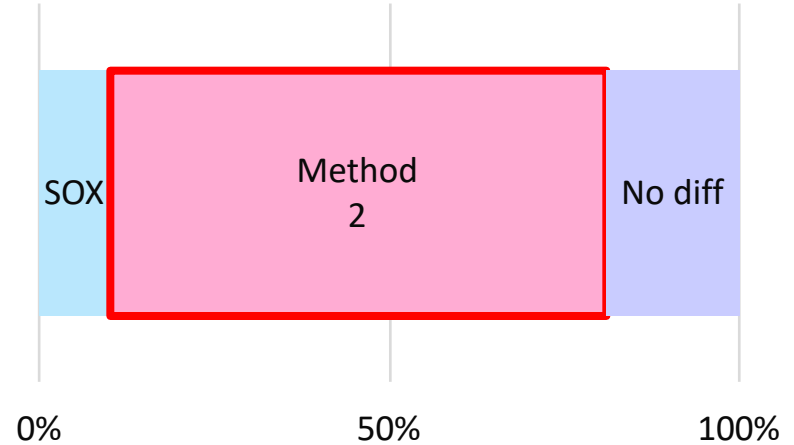
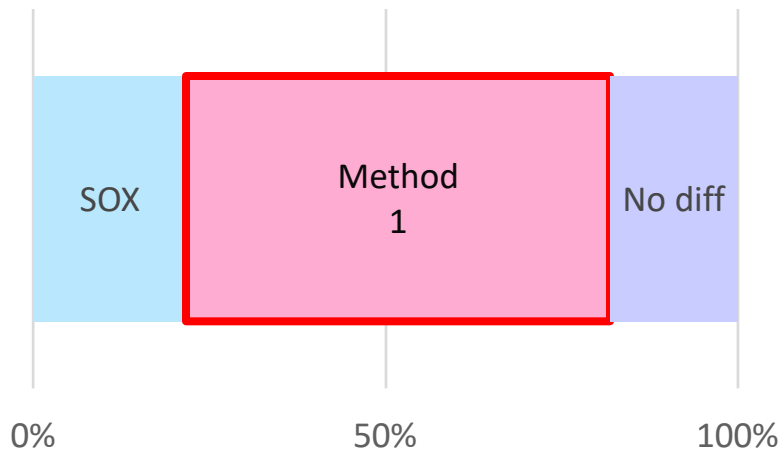
	fast	normal	slow
Proposed Method 1 (phonetic symbol expansion)			
Proposed Method 2 (add ratio of speaking rate)			
Speaking rate changed through post-processing (using Sox)			

Experiments 1: Methods

- Subjects were randomly presented with two utterances and asked which sounded more natural



Experiments 1: Results






- Proposed methods are perceived as more natural than post-processing
- No difference in quality between the proposed methods

Experiments 2: Methods

- **Effectiveness of the phoneme level speaking rate variation**

- Does the proposed method change the speaking rate at the phoneme level?

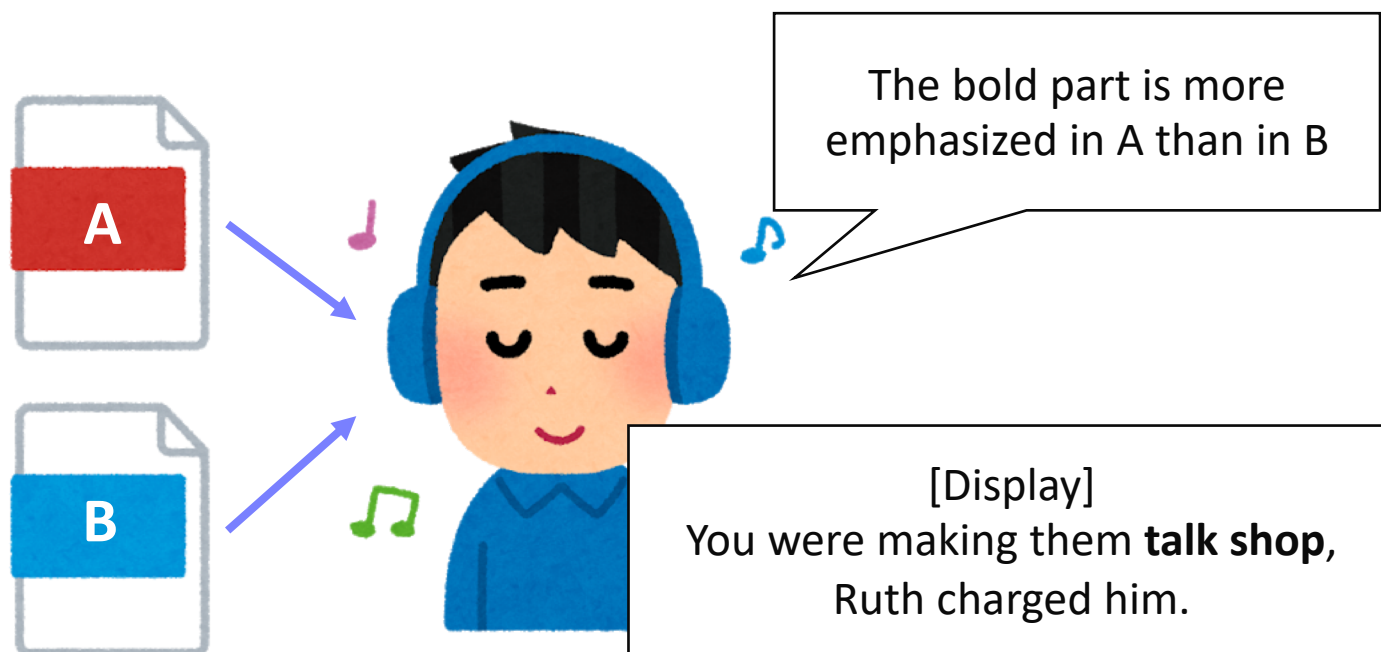
- Generated 3 types of speech

[Normal] “You were making them talk shop, Ruth charged him.”	
[fast] “You were making them talk shop, Ruth charged him.”	
[proposed] “You were making them talk shop , Ruth charged him.”	

normal: black, fast: blue, slow: red

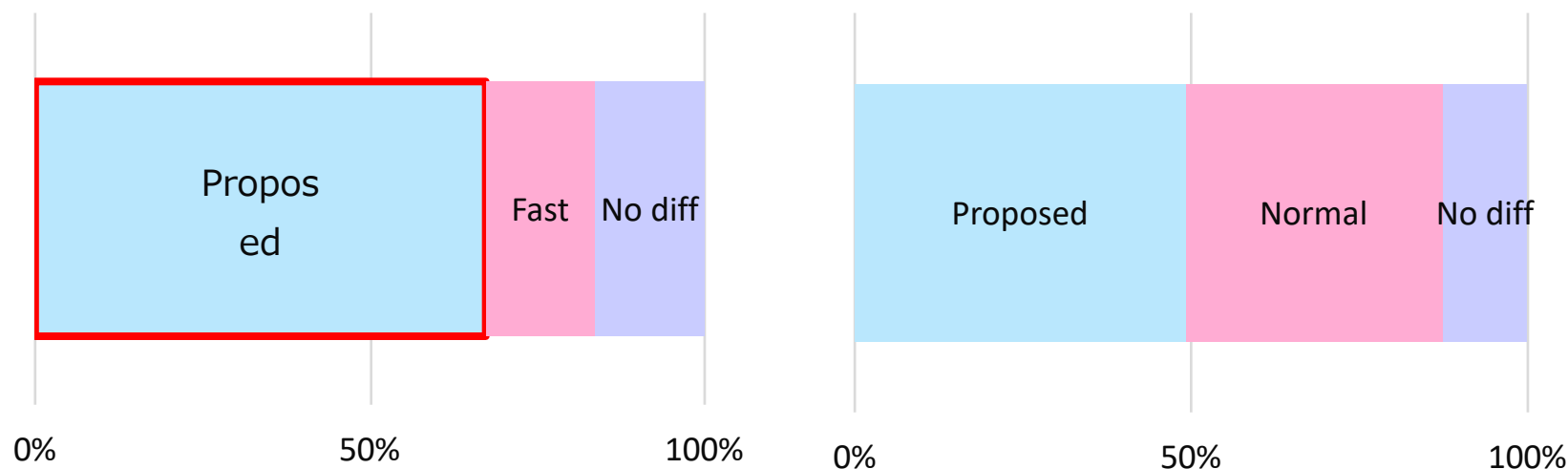
Experiments 2: Methods

- Subjects were presented with text containing emphasis markers on certain words (bold print), and two randomly selected utterances
- They were asked which sample featured stronger emphasis



Experiments 2: Results

- The results showed that the proposed method can change the speaking rate appropriately.
- There was no difference between the proposed method and the “normal” speech
 - The recorded “normal” speech was slower than regular natural reading speech



Conclusion

- **GAN-TTS that controls the speaking rate variation at the phoneme level**
- **Methods**
 - Phonetic symbol expansion
 - Add ratio of speaking rate
- **Experiments**
 - Subjective assessment of the naturalness of speech
 - Effectiveness of the phoneme level speaking rate variation
- **Results**
 - The proposed method was perceived as more natural than manipulating the waveform of synthetic speech with post-processing
 - It successfully performed speaking rate variation at the phoneme level