Phonetic and Prosodic Features for Sequence-to-Sequence Acoustic Modeling on Japanese Text-to-Speech and Their Estimation

March 2024

Kiyoshi Kurihara

Phonetic and Prosodic Features for Sequence-to-Sequence Acoustic Modeling on Japanese Text-to-Speech and Their Estimation

Graduate School of Science and Technology

Degree Programs in Systems and Information Engineering

University of Tsukuba

March 2024

Kiyoshi Kurihara

Abstract

Thanks to the rapid progress in deep learning technology, the text-to-speech (TTS) method has achieved the same quality as human speech. In this thesis, we propose a method for text-to-speech synthesis and estimation that utilizes phonetic and prosodic features. And, we propose a method for speech synthesis that allows for the adjustment of acoustic features.

In 2023, speech synthesis technology will be in practical use in many applications. Throughout its evolution, speech synthesis technology has progressed from deep neural network (DNN)-based statistical speech synthesis to sequence-to-sequence (seq2seq) with attention-based text-to-speech (TTS). Notably, seq2seq with attention-based TTS reached subjective evaluation equivalence with human-recorded speech following the introduction of WaveNet, a neural network-based speech generation technology. However, challenges arose in implementing seq2seq-based speech synthesis for Japanese. The issue stemmed from the fact that Japanese character strings possess multiple readings, causing a mismatch between the strings and speech, leading to suboptimal learning outcomes.

This thesis proposes a method for Japanese speech synthesis utilizing seq2seq with attention-based TTS which using phonetic and prosodic features. The incorporation of reading phonetic and prosodic features as high/low pitch accent controls, corresponding to pitch-accented languages, is crucial for achieving accurate Japanese pronunciation. This paper introduces a method for Japanese speech synthesis incorporating the phonetic and prosodic features, along with a technique for inferring phonetic and prosodic features from speech. We also propose a method capable of estimating phonetic and prosodic features directly from speech. This approach enables the estimation of phonetic and prosodic features based solely on speech input, facilitating the generation of labels for speech synthesis. Additionally, we propose a method for adjusting acoustic features in speech synthesis.

Acknowledgments

I would like to express my deepest gratitude to all those who supported and advised me during the research activities related to this thesis.

Firstly, I extend my sincere appreciation to my advisor, Prof. Takeshi Yamada at the University of Tsukuba. He kindly accepted my request to supervise me, even though I was a complete stranger. Despite the limited duration of one year, he provided me with significant support.

I also want to express our gratitude to the NHK members who collaborated with me in our research activities. Special thanks to Ms. Mayumi Abe and Mr. Tetsushi Okura, who, alongside me, received the Award for Technological Advancement and Development (Field Operation Category) from The Institute of Image Information and Television Engineers, as well as the Broadcasting Technology Division Award from the 49th Hoso Bunka Foundation. I am thankful for the invaluable support they offered in the practical implementation of text-to-speech technology.

Additionally, I express our gratitude to Mr. Nobumasa Seiyama of the NHK Foundation, who worked with me in our research activities. He provided valuable guidance on text-to-speech during his tenure at NHK Science & Technology Research Laboratories, supporting the ideas presented in this thesis and contributing to the preparation of the paper.

I extend my thanks to all the members of the multimedia laboratory at the University of Tsukuba. Despite having much non-research-related duties at the company, I found solace in listening to their research progress during our weekly speech team meetings.

Lastly, I want to express my heartfelt thanks to my family for their unwavering support and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Background

Thanks to the rapid progress in deep learning technology, the text-to-speech (TTS) method has achieved the same quality as human speech. Deep neural network (DNN)-based statistical speech synthesis [1], which emerged in 2013, significantly improved the quality of synthesized speech. Additionally, sequence-to-sequence (seq2seq) with attention-based TTS [2], now a mainstream approach, was introduced in 2017. Yet, when seq2seq speech synthesis first emerged, no method had been introduced to support Japanese. The purpose of this thesis is to explore phonetic and prosodic features (PPF) in conjunction with seq2seq acoustic modeling, with the aim of achieving seq2seq-based Japanese speech synthesis and estimating PPFs from speech. Moreover, there currently exists no DNN-based statistical speech synthesis system capable of dynamically adjusting speaking style, pitch, and speech rate. Furthermore, there is a lack of versatile speech synthesis solutions suitable for a wide range of content.

In this thesis, we propose three methods to address this issue. First, the seq2seq models for speech synthesis [2] delivers high-quality speech but does not accommodate Japanese pitch accents. Second, we propose the novel method for estimating Japanese prosodic features in speech. There was no method for realizing phonetic and prosodic labels estimation from speech [3]. Third, we propose a method for analytically adjusting acoustic features within the intermediate process of the DNN statistical parametric speech synthesis pipeline [1] to vary the acoustic features and to learn various speaking styles [4] and speakers in a single model.

1.2. Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS

To achieve high-quality speech synthesis in Japanese that matches the sound quality of recorded human speech, it was necessary to develop sequence-to-sequence with attention (seq2seq)-based TTS specifically for Japanese [5]. In 2017, speech synthesis using the seq2seq models emerged [6]. This method demonstrated performance similar to end-to-end speech synthesis in English. However, unlike English, Japanese characters encompass various types of kanji, katakana, and hiragana. With multiple readings associated with kanji characters, there is no direct correspondence between character strings and their phonemes. Consequently, the mismatch between character features and acoustic features prevents successful training [7]. In addition, katakana alone has the problem of irregularly changing accents. To address this issue, we propose a novel prosodic features control method for Japanese speech and develop a speech synthesis that can leverage the seq2seq-based TTS for Japanese. This advancement has allowed us to achieve high-quality speech synthesis.

1.3. Phonemes and Prosodic Feature Recognition

Japanese is a pitch accent language and accent rules operates on binary information [8]. This accentual rule plays a crucial role in understanding how speech is delivered in Japanese. For instance, similar to Japanese diacritics, pitch accent information is utilized in Japanese accent dictionaries [9], which are indispensable tools for the professional anchors, voice actors and Japanese learners. The conventional approach to estimating accentual information through machine learning from the F0 of the speech signal [10] often yields a low recognition rate and is unsuitable for training data in speech synthesis. But our novel phonetic and prosodic labels data are proprietary and available in limited quantities data. Therefore, it was necessary to devise methods to improve recognition accuracy with a small amount of data. In response to this challenge, we propose a method for estimating labels using the self-supervised learning acoustic modeling (AM) estimation approach [3]. This method can be effectively trained even with a limited amount of data.

1.4. Speech Synthesis with Adjustable Acoustic Features

DNN-based statistical parametric speech synthesis (SPSS) [1] was limited in its ability to adjust speaking style, speech rate, pitch, and intonation. However, audio contents require various types of speech features, and professional anchors have access to it, thereby enhancing the content's quality. In 2018, Hojo proposed a speech synthesis method that can switch speakers [11]. However, a method for controlling speaking style has not been proposed. To address these limitations and enable flexible adjustments of speech characteristics, we have a proposed method capable of modifying speaking style, speaker, speech rate, pitch, and intonation [12]. While there are various potential applications, our focus lies on two methods. The first method involves training a corpus with multiple speaking styles and speakers in a single model, controlling for them to enhance overall quality. The second involves adjusting the intermediate acoustic features of the DNN-based SPSS [12] to correct speech rate, pitch, and intonation.

1.5. Overview of Thesis

This thesis consists of four parts. In Chapter 2, we propose a Japanese text-to-speech method that utilizes the seq2seq model which PP labels [5] to accurately represent speech through symbols. Additionally, we discuss the evaluation metrics employed in this study. In Chapter 3, we propose phonetic and prosodic labels estimation method for speech [3]. Finally, in Chapter 4, we propose the concept of speech synthesis with speaking styles and speakers adjustable acoustic features [4].

Chapter 2

Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS

2.1. Introduction

One of the seq2seq-based TTS methods, also known as Tacotron 2 [2] has achieved speech quality on par with human speech in English. The conventional TTS method, known as DNN-based SPSS [1], was only able to develop a neural network for acoustic features, while other components relied on analytical methods, resulting in low sound quality. Consequently, we decided to develop neural networks for all components and initiate research on Japanese speech synthesis using seq2seq [5] to improve the synthesized speech. However, we encountered difficulties in training seq2seq TTS for Japanese due to the high/low pitch accent and the presence of kanji with over two thousand types of characters and multiple readings. Japanese consists of three types of characters: Kanji, hiragana, and katakana. Specifically, there are 2,136 commonly used kanji characters [13], and over 30% of them have multiple readings, making them challenging to train. The emergence of seq2seq-based TTS allowed the training data to transition from specially formatted data for speech synthesis to human-use character strings. However, even with seq2seq-based TTS, training kanji characters presented issues due to a large variety of kanji characters and multiple readings. Of the 2,136 kanji characters in the regular use set, slightly more

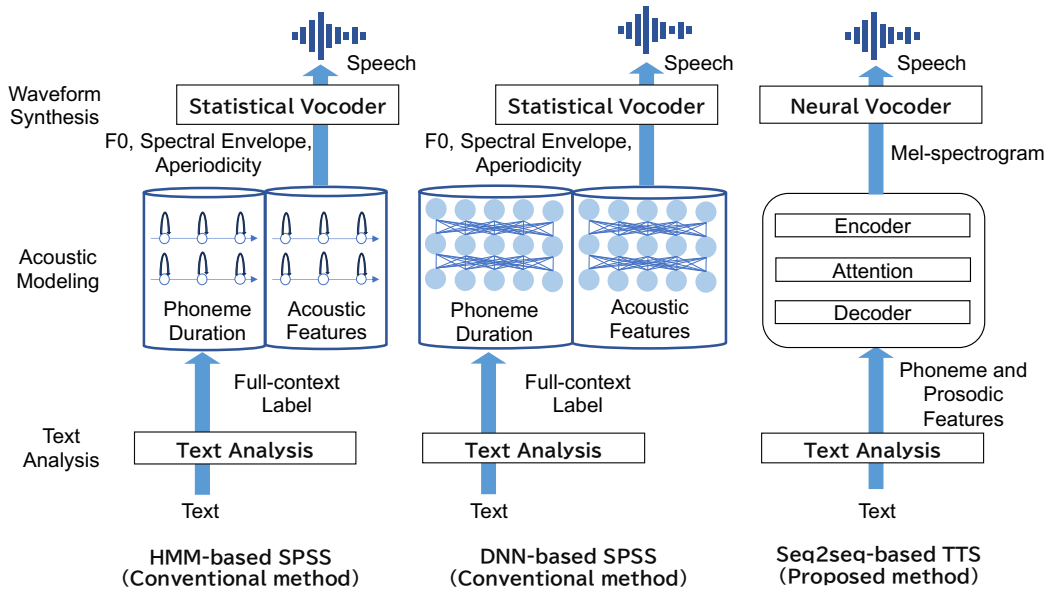


Figure 2.1: Transition of Speech Synthesis Method.

than 30% have multiple readings. Seq2seq-based TTS is a method for matching strings and acoustic features. Consequently, when there are numerous patterns in which strings and acoustic features do not align, the training process becomes challenging. On the other hand, katakana, a common way of representing Japanese readings, does not provide information about high/low pitch accents to indicate the correct reading. In seq2seq, English stress accents have been confirmed to be trained correctly [2], but a notation for representing high/low pitch accents used in Japanese has not been proposed. We propose a method to utilize the Dictionary of Japanese Pronunciation and Accentuation [9] as input for seq2seq-based TTS, employing katakana and prosodic features to represent Japanese speech. We have confirmed that the phonetic and prosodic labels (PP labels) can be applied effectively to various types of seq2seq-based TTS. Since the introduction of seq2seq-based TTS in 2017, there has been no method of speech synthesis capable of directly training raw Japanese character strings, even as of 2023.

2.2. Conventional Method and Its Problems

The conventional methods contain two types of TTS method (Figure 2.1). First is hidden Markov model (HMM)-based statistical parametric speech synthesis [14], second is DNN-

based SPSS. DNN-based TTS [15] was developed using HMM-based statistical parametric speech synthesis [14]. Training data of this method, full-context (FC), contains many features and have low readability. It also requires manually corrected phoneme alignment, which is expensive to work with. End-to-end speech synthesis has been actively researched ever since Tacotron 2 [16] first produced English speech comparable in quality to that of human speech. Tacotron 2 uses a seq2seq model, as do similar methods such as Tacotron [17], Char2Wav [6], VoiceLoop [18], Deep Voice 3 [19], and Transformer-based TTS [20].

In related research, seq2seq AM with a dependency on accent input is proposed. The system of Yasuda [7] considered input pitch-accent information in Tacotron, but the system controlled only accent information, and the quality of the synthesized speech was worse than that of conventional SPSS. Moreover, it was only compatible with Tacotron. Yasuda's method inputs a phoneme sequence and accentual-type sequence separately and embeds them separately. Our method differs from Yasuda's in terms of the accent sequence format. In our method, the PP labels are merged into one sequence and thus more readable than their method that has to separately input sequences. In addition, Fujimoto [21] conducted an experiment comparing the suitability of phoneme and mora as units of the input sequence, which are one-hot vectors or linguistic features, of seq2seq AM. The sequence in the phoneme method is about twice as long as that in the mora method, but long sequences tend to occur alignment error [22]. There is a possibility that the phoneme method causes more alignment errors than the mora method even though there is no difference between phoneme and mora in terms of pronunciation. The system of Okamoto [23] inputted FC labels [24] to Tacotron 2, but it could not input symbols directly and was only compatible with FC labels that have poor readability. Although seq2seq AM can handle symbols directly, the system of Okamoto does not take advantage of the benefits of simple input symbols. Furthermore, it is only compatible with Tacotron 2. Shechtman [25] proposed a method that reproduces prosody information. The attention section in the seq2seq AM of this method has a recurrent architecture. The method produces expressive speech and can control the duration, but it cannot replicate prosodic features. While the present study is related to these recent approaches in seq2seq AM, it is compatible with

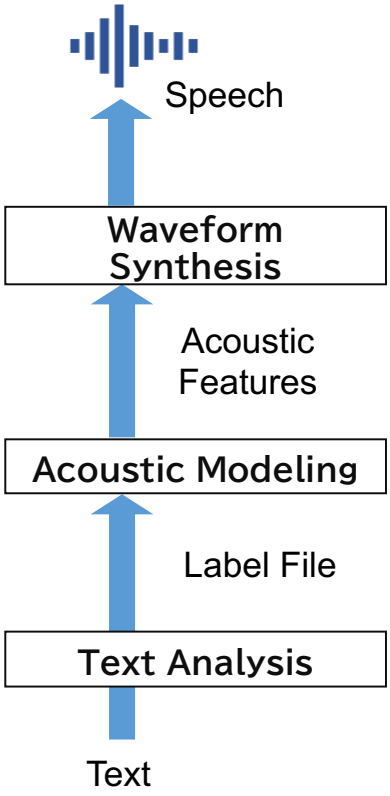


Figure 2.2: Overview of Common Text-to-Speech Method.

multiple seq2seq AM methods, controls prosodic features, and has directly readable and writable data descriptions, aspects that were not considered in the earlier studies.

2.2.1. Overview of Text-to-Speech Synthesis

The TTS can be broadly divided into three sections (Figure 2.2). The first is “text analysis,” the second is “acoustic modeling” and the third is “waveform synthesis.” The text analysis section estimates the type of sequences generated from the text, making it the section that determines pronunciation, commonly referred to as grapheme-to-phoneme (G2P) [26]. It mainly uses natural language processing (NLP) technology for this purpose. The acoustic modeling section estimates acoustic features by inputting specific strings and affects sound quality as a result. This section has undergone significant improvements in performance due to contributions from deep learning. Training data of conventional SPSS is FC labels [24]. Acoustic features of SPSS estimate duration time (time alignment) of phoneme and

vocoder parameter consist of fundamental frequency (F0), spectral envelope and aperiodicity. In many cases, SPSS consist of feedforward DNN or long short-term memory (LSTM) [27] waveform synthesis section transforms acoustic features to waveforms. This part is referred to as a vocoder also known as “statistical vocoder [28]” or “neural vocoder [29].”

2.2.2. Text-to-Speech Synthesis

With the introduction of HMM-based TTS, the capability of generating speech for arbitrary sentences became a reality in speech synthesis. However, HMMs exhibited low accuracy. The advent of DNN-based TTS marked a significant improvement in quality. Both methods supported input formats using FC labels and demonstrated proficiency in training and inference within specific formats. The appearance of a waveform generation method called WaveNet [30] in 2016 and that of the seq2seq method [2] [6] of TTS in 2017 brought the quality of TTS to a level equivalent to that of human speech. In addition, the training data became simpler and less expensive to produce. In the context of seq2seq-based TTS, the input string is a sequence, the nature of which varies based on the language. For languages like English, where strings and pronunciations closely align, seq2seq-based TTS exhibited smooth training. However, for pitch accent languages such as Japanese, where a single character can have multiple readings, as in the case of kanji, the need to devise a specific string for training arose. In the instance of a pitch accent language like Japanese, a tailored learning string was imperative.

2.2.3. Text Analysis

The text analysis section consists of the method that converts graphemes of text spelled out into phonemes, which is why it is called a grapheme-to-phoneme (G2P) method [26]. If not done well, this estimation of phonemes will negatively affect pronunciation, resulting in strange intonation. If the intonation of synthesized speech is felt to be unnatural, the reason for this is thought to be the quality of G2P. The G2P method differs from one language to another. In the case of English, many words can be straightforwardly converted because many of them are one-to-one conversion in graphemes and phonemes, but because some words have multiple readings, using deep learning methods has been proposed [26].

We propose novel G2P method for acoustic modeling of pitch-accent language [5].

2.2.4. Acoustic Features Estimation and Speech Generation

Waveform audio file consist of 48k discrete data points per second in CD quality, resulting in a substantial amount of data for generation. Consequently, the task of speech synthesis, which aims to produce accurate speech data from a limited representation of linguistic feature, is often divided into “acoustic features estimation (acoustic modeling)” and “speech generation (waveform synthesis).” The acoustic features are estimated from linguistic features obtained through text analysis. In DNN-based SPSS, DNNs are trained to estimate acoustic features from FC labels. Although this acoustic feature information is less comprehensive than raw data speech, it retains the linguistic information of the speech and is used in speech visualization methods.

The task of accurately generating waveforms is challenging due to the limited amount of data available for acoustic features and the significantly larger amount of data required for waveforms. To address this issue, various methods have been proposed, including analytically calculating waveforms from acoustic features such as F0, spectral envelope, and aperiodicity. Additionally, methods involving the estimation of waveforms from mel-spectrograms using convolutional neural networks (CNNs) [31] have been introduced.

2.3. Proposed Method

We develop a method of controlling the prosodic features that works by inputting a symbol between phonetic symbols. Seq2seq AM corresponds to inputs of phonetic symbols [17], but it has been confirmed that the accents cannot be reproduced by training input consisting of Japanese phonetic symbols [32] due to the lack of accent information. Furthermore, normal Japanese text contains multiple readings of characters, which makes it unsuitable for seq2seq AM training. To solve these problems, we propose a notation method that represents the synthesized speech uniquely by using PP labels. The prosodic symbols (Table 2.1) complement the acoustic features between phonetic symbols and can reproduce accurate speech. We conducted experiments show that the proposed method is effective for pitch-accent languages such as Japanese.

PP labels can be generated automatically and using them to make annotations is simple. The conventional DNN-based SPSS [1] [30] uses time-aligned FC labels. A high-quality SPSS method requires accurate time alignments, and the cost of accurately determining time alignments is high. Our method does not require time alignment because the seq2seq-based TTS can train directly from the input symbols. We propose a way of automatically converting FC labels into PP labels. FC labels contain linguistic features such as phoneme, accent information, accentual phrase boundary, end-of-sentence (EOS), and pause. This information is rearranged according to the proposed rules, and the results are natural synthesized speech in seq2seq AM. Evaluations have shown that using PP labels yields more natural speech than conventional SPSS. This shows the possibility of their general use with seq2seq AM methods. We conducted objective and subjective experiments, and the results indicated that placing prosodic symbols between phonetic symbols can control accents, pause breaks, accentual phrase boundary, and EOS. We confirmed that PP labels can control prosodic features by using three seq2seq AM methods in TTS for Japanese.

2.3.1. Sequence-to-Sequence Acoustic Modeling for Neural TTS

Seq2seq AM is a method that generates mel-spectrograms representing inputted symbols.

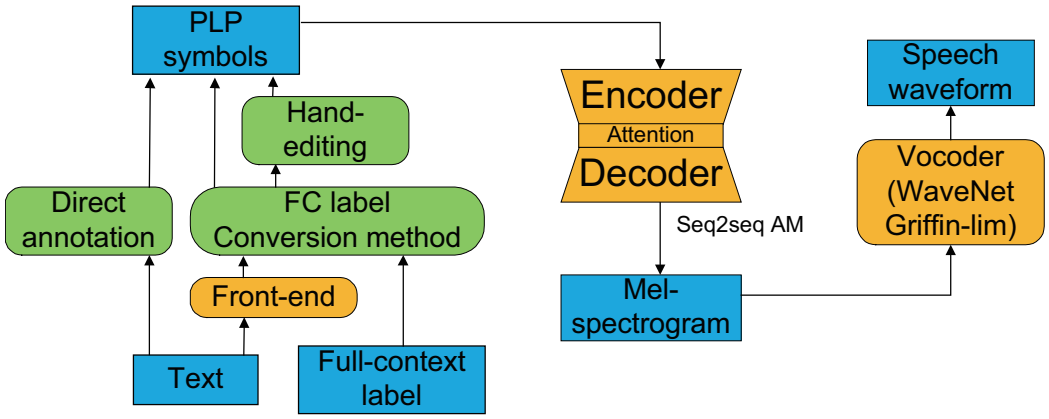


Figure 2.3: Diagram of TTS Using Seq2seq AM and PLP Labels.
（Copyright (C) 2021 IEICE, [5] Fig. 1）

Since the advent of WaveNet [30], it has enabled high-quality speech to be generated from mel-spectrograms. The advantage of this approach is that mel-spectrogram can be converted directly into a waveform. To reproduce prosodic-acoustic features such as accent for seq2seq AM, we insert prosodic symbols between phonetic symbols and use the result as the input of seq2seq AM. This method represents the synthesized speech uniquely, especially pitch accents, and should be easily readable by people. It is highly readable, so the labels can be read and understood directly, and a front-end is not required, depending on how it is used.

The right side of Figure 2.3 shows an overview of a TTS system incorporating the proposed method. This diagram consists of three parts. The first part is the PP label generation. By combining the FC label conversion method with the front-end, the PP labels can be automatically generated from the text. In addition, the FC labels can be used and the information in the existing labels, can be discarded. The second part is the mel-spectrogram estimation with seq2seq AM. This part can be replaced with other seq2seq architectures. The third part is waveform generation methods convert acoustic features, such as mel-spectrograms, into waveforms, which are discrete signals. WaveNet, which uses CNN, and Griffin-Lim vocoder, which estimates phase from acoustic features, are common methods. Recently, WaveNet is computationally expensive, so faster methods using generative adversarial network (GAN) [33] and generative flow (GLOW) [34] are

Table 2.1: Examples of Prosodic Symbols.

Feature	Prosodic symbols
Initial rising	^
Accent nucleus	!
Accentual phrase boundary	#
EOS (Declarative)	(
EOS (Interrogative)	?
Pause	_

commonly used.

2.3.2. Prosodic Symbols

Table 2.1 lists the PP labels. We specified the prosodic symbols with reference to the Japanese ToBI label model [35] as follows. The prosodic symbols consist of initial rising (which denotes a rapid rising of F0 after the symbol), accent nucleus (which denotes a rapid falling of F0 after the symbol), accentual phrase boundary, EOS, and pause symbols. An accent-phrase is a unit that forms an accent during pronunciation. This description method uses arbitrary symbols, but has a simple representation. Because the symbols are arbitrary, this method can be applied to other languages besides Japanese.

2.3.3. English and Japanese Accent System and Our Notation Method

Tacotron 2 employs the ARPAbet [36] notation for training English. ARPAbet represents phonemes and allophones of General American English with distinct sequences of ASCII characters. The conventional method for converting English text to ARPAbet is through G2P language processing in English. In Japanese, the G2P approach is used, as shown in Chapter 2.1. The conventional method for Japanese speech production involves a seq2seq method with a dependency component. Meanwhile, we propose a versatile Japanese

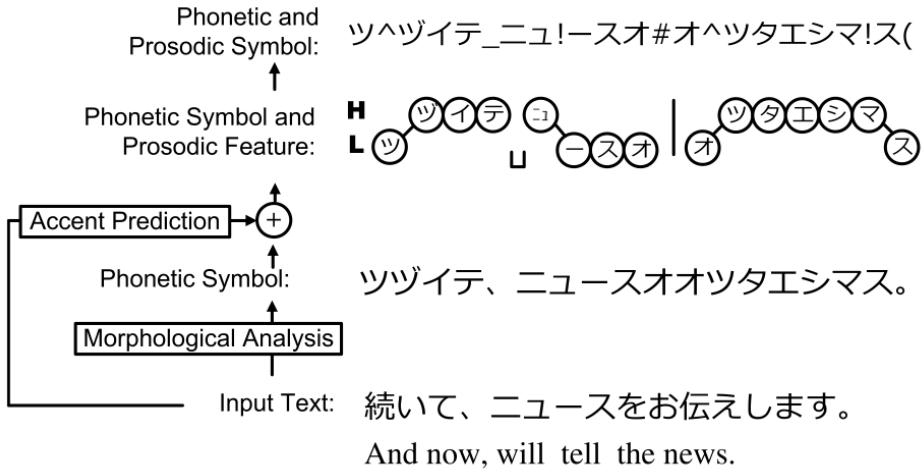


Figure 2.4: Procedure of Producing Phonetic and Prosodic Symbols.
（Copyright (C) 2021 IEICE, [5] Fig. 2）

language method that is based on PP labels as input.

Figure 2.4 illustrates our notation method for representing high (**H**) and low (**L**) pitch between mora for seq2seq AM. The Japanese language has a pitch-accent feature that can be represented as a sequence of binary F0 levels in mora units [37] [38]. Japanese pitch accents in Tokyo dialect have the following rules.

• A rapid rise or fall in F0 must take place between the first mora and the second mora.

• The maximum number of rapidly rising patterns of F0 between two consecutive morae in a word is one.

• The maximum number of rapidly falling patterns of F0 between two consecutive morae in a word is one.

We have added these rules to the notation method for seq2seq AM and made up the label format for seq2seq AM. Figure 2.4 shows the procedure of producing phonetic and prosodic symbols. First, we conduct morphological analysis [39] for obtaining PP labels to represent phonetic symbols and accent prediction [40] for obtaining accent information. Second, we combine these sequences and the representations for the phonetic symbols and prosodic features. Finally, we follow the accent rules in Table 2.1; we place the prosodic symbols between the phonetic symbols and obtain phonetic and prosodic symbols. In the

inputted PP labels, the initial rising "^" and accent nucleus "!" prosodic symbols make the prosodic-acoustic feature appropriate high or low pitch until the next prosodic symbol appears in the input of seq2seq AM.

2.3.4. Text Analysis and FC-Label-to-PP-Label Conversion

Open JTalk [40] is a TTS system that contains a Japanese front-end, and it converts Japanese text into FC label [24]. Open JTalk contains Japanese text analysis functions, including grapheme-to-phoneme conversion, and a morphological analysis called Mecab [39]. Figures 2.5 and 2.6, and Tables 2.2 and 2.3 explain the FC label converter to PP label. The FC label has linguistic and acoustic information; we pick up the phonemes and features listed in Table 2.1 and convert the sequences into PP symbols. On the other hand, Japanese kanji characters potentially contain mismatches between input symbols and acoustic features, which can potentially cause alignment errors in the encoder outputs and decoder inputs due to the kanji having multiple readings. These mismatches may in turn cause alignment errors depending on conversion errors in the text of the corpus.

Algorithm: Full-context label conversion method

Input: Full-context label, N=number of phonemes
Output: Phonetic and prosodic symbols
for n ← 1 to N do
  $PP_n \leftarrow PP_{n-1}.append(p_{n,3})$,
  if $a_{n,3}$=1 **and** $a_{n+1,2}$=1 **then**
    $s_n \leftarrow$ "#", $PP_n \leftarrow PP_n.append(s_n)$,
  else if $a_{n,1}$=0 **and** $a_{n,2}$!=$f_{n,1}$ **then**
    $s_n \leftarrow$ "!", $PP_n \leftarrow PP_n.append(s_n)$,
  else if $a_{n,2}$=1 **and** $a_{n+1,2}$=2 **then**
    $s_n \leftarrow$ "^", $PP_n \leftarrow PP_n.append(s_n)$,
  else if $p_{n,3}$="pau" **then**
    $s_n \leftarrow$ "_", $PP_n \leftarrow PP_n[1:-2].append(s_n)$,
  else if $P_{n,3}$ = "sil" **and** $n = N$ **then**
    if $e_{n,3}$ =0 **then**
      $s_n \leftarrow$ "(", $PP_n \leftarrow PP_n.append(s_n)$,
    else if $e_{n,3}$ =1 **then**
      $s_n \leftarrow$ "?", $PP_n \leftarrow PP_n.append(s_n)$,
end

Figure 2.5: Algorithm of Full-Context Label Converter.
（Copyright (C) 2021 IEICE, [5] Fig. 3）

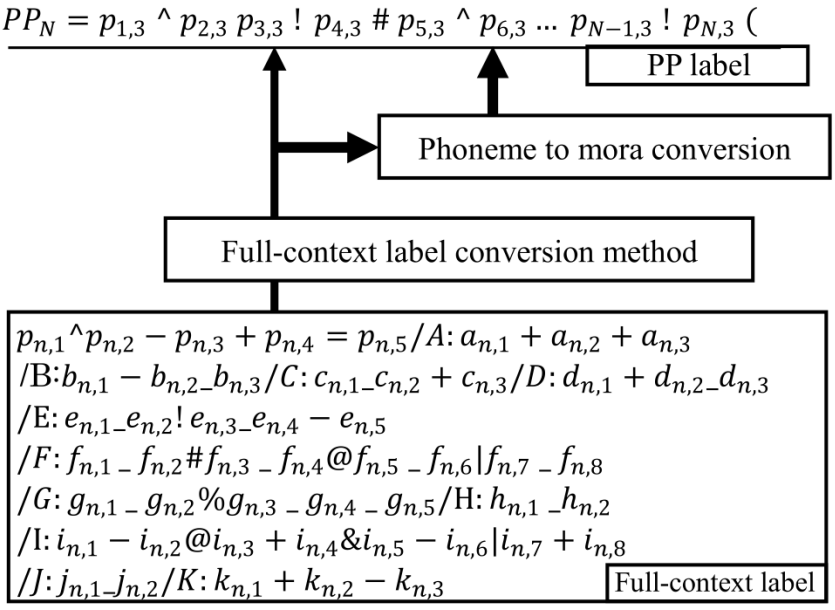


Figure 2.6: Overview of Converting Full-Context Label into Phonetic and Prosodic Symbols.
（Copyright (C) 2021 IEICE, [5] Fig. 4）

Table 2.2: List of Context Feature Templates.

Index	Feature templates
n	Order of phoneme symbol
p_n	Phoneme identifies
a_n	Accent type and position
b_n	Part-of-speech, inflected and conjugation of previous word
c_n	Part-of-speech, inflected and conjugation of current word
d_n	Part-of-speech, inflected and conjugation of next word
e_n	Information on previous accent phrase
f_n	Information on current accent phrase
g_n	Information on next accent phrase
h_n	Information on previous breath group
i_n	Information on current breath group
j_n	Information on next breath group
k_n	Number of breath groups, accent phrases and moras
PP_n	Symbols of phonetic and prosodic features

Table 2.3: List of Context Feature Templates.

Index	Feature templates
$p_{n,3}$	The current phoneme identity
$a_{n,1}$	The difference between accent type and position of the current mora identity
$a_{n,2}$	Position of the current mora identity in the current accent phrase (forward)
$a_{n,3}$	Position of the current mora identity in the current accent phrase (backward)
$e_{n,3}$	Whether the previous accent phrase interrogative or not (0: not interrogative, 1: interrogative)
$f_{n,1}$	The number of moras in the current accent phrase
pau	Information on pause
sil	Information on silence

2.3.5. Hand-Editing and Direct Description

The PP labels are simple and can be edited by hand. This enables direct annotation without an FC label converter. Compared with time-aligned FC labels, PP labels do not require the boundary positions of the phonemes, the identification of which is a time-consuming task because the boundaries between phonemes are ambiguous. In particular, with PP labels, we can read and annotate them directly. Meanwhile, it is not possible to read FC labels because of their complicated expression.

2.4. Experiments

We conducted objective evaluations on the encoder-decoder alignments, synthesized mel-spectrograms and F0. We also conducted subjective evaluations of the naturalness of the synthesized speech.

2.4.1. Datasets and Experimental Conditions

We used the JSUT [41] corpus, which is a large-scale open Japanese speech corpus. The whole corpus contains 10 hours of speech and corresponding normal Japanese text. It contains 7,696 utterances. The corpus was split into 7,596 samples for training and 100 samples for testing. The test set included 30 samples for evaluations and four samples for the evaluators' training. FC labels with time alignments were generated by Open JTalk and Julius [42] [43] and PP labels were generated by Open JTalk and the FC label conversion method (see Section 2.3.4). The number of training iterations for the WaveNet vocoder [29] was 860,000. The sampling rate was 22 kHz, 16-bit.

We conducted experiments comparing four types of input sequences consisting of commonly used Japanese characters and three types of seq2seq AM methods and implementations. We automatically prepared the Japanese PP labels by using the method described in Section 2.3.4. The first type was normal Japanese text consisting of kanji and hiragana (KH). Hiragana has the same reading as katakana. The second type was automatically generated plain katakana (KT) which represented readings without accent information; the characters were represented by mora. The third type was phonetic symbols

consisting of Roman alphabet (phoneme) and prosodic symbols (PP (phon.)). The fourth type was phonetic symbols consisting of katakana (mora) and prosodic symbols (PP (mora)). The PP labels and KT included misread kanji converted by Open JTalk, but they were used in the training without making any hand-edited corrections to them. In Sections 2.4.3 and 2.4.6, we also used the PP (phon.) for comparison with PP (mora). These types of sequences composed the input of the proposed method. Table 2.4 lists the systems that were tested in the experiment, and Table 2.5 shows an example of input symbols.

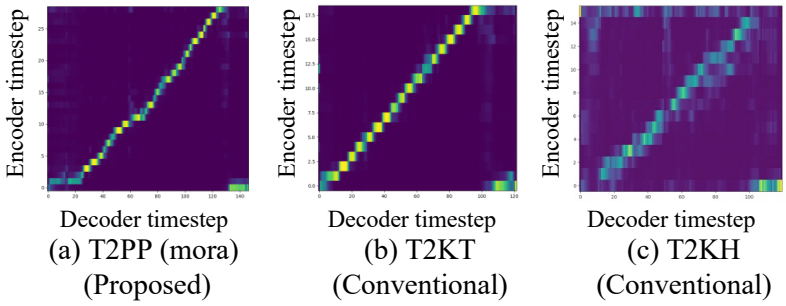
Mel-spectrograms generated from seq2seq AM were converted into waveforms by using WaveNet or a Griffin-Lim vocoder. The mel-spectrogram features of all methods had 80 dimensions, a 125 - 7600-Hz frequency band, and a 46 ms window size.

Table 2.4: Systems Compared in the Experiments.

System	AM method and implementation	Input feature
T2KH	Tacotron 2 [43]	KH
T2KT	Tacotron 2 [43]	KT
T2PP (phon.)	Tacotron 2 [43]	PP (phon.)
T2PP (mora)	Tacotron 2 [43]	PP (mora)
DV3KH	Deep Voice 3 [42]	KH
DV3KT	Deep Voice 3 [42]	KT
DV3PP	Deep Voice 3 [42]	PP (mora)
TRKH	Transformer-based TTS [20]	KH
TRKT	Transformer-based TTS [20]	KT
TRPP	Transformer-based TTS [20]	PP (mora)

Table 2.5: Example of Input Symbols.

Input features	Example of input symbols	Number of symbols
KH	私の席は、あの婦人の横ですか。	15
KT	ワタシノセキワ、アノフジンノヨコデスカ。	20
PP (phon.)	wa^tashino#se!kiwa_a^no#fu^jiNno#yo^kodesu!ka(	46
PP (mora)	ワ^タシノ#セ!キワ_ア^ノ#フ^ジンノ#ヨ^コデス!カ(	29



(a) T2PP (mora) (Proposed) (b) T2KT (Conventional) (c) T2KH (Conventional)

Figure 2.7: Alignment of Encoder and Decoder.
（Copyright (C) 2021 IEICE, [5] Fig. 5）

2.4.2. Visualization of the Encoder-Decoder Alignment

Figure 2.7 shows the alignments of T2PP (mora), T2KT, and T2KH. The iterations and batch size of Tacotron 2 were 600,000 and 48, respectively. Figure 2.7 (a) shows a slightly non- smooth alignment. In this case, the combination of a phonetic and a prosodic symbol expressed an acoustic information of a mora as decoder timesteps.

The figure indicates that the alignment monotonically increased and was continuous. Figure 2.7 (b) shows a smoother alignment; there was a one-on-one correspondence between the encoder which represented a mora character and the decoder timestep which represented the mora mel-spectrogram. Figure 2.7 (c) shows unclear and discontinuous alignments in some timesteps, suggesting that T2KH had difficulty training the model.

2.4.3. Objective Evaluation of Prosodic Symbols

Figures 2.8 and 2.9 compare the mel-spectrograms and F0 of T2PP (mora), T2KT, and T2KH. The iterations and batch size of Tacotron 2 were 600,000 and 48, respectively. The proposed method (T2PP (mora)) was better at reconstructing the details in the red rectangles, and they replicated pauses and the falling and rising of tone and pitch. The ground truth and T2PP (mora) had similar features both overall and in detail. In contrast, T2KT produced overall flat speech; it did not reproduce the pitch- accents and pauses in the red rectangles. The T2KH was not similar in shape to the ground truth.

Moreover, the results in Figures 2.8 and 2.9 confirmed that the accent nucleus “!”

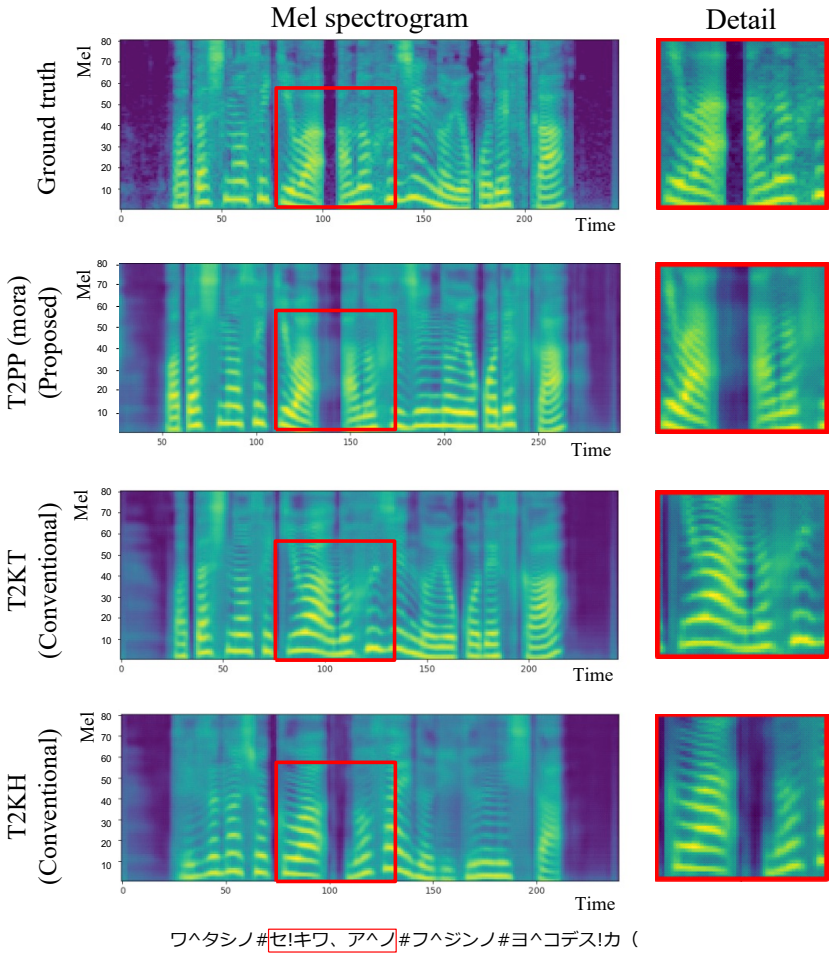


Figure 2.8: Comparison of Mel-Spectrograms for “私の席は、あの婦人の横ですか。 (Is My Seat Next to That Lady?).”

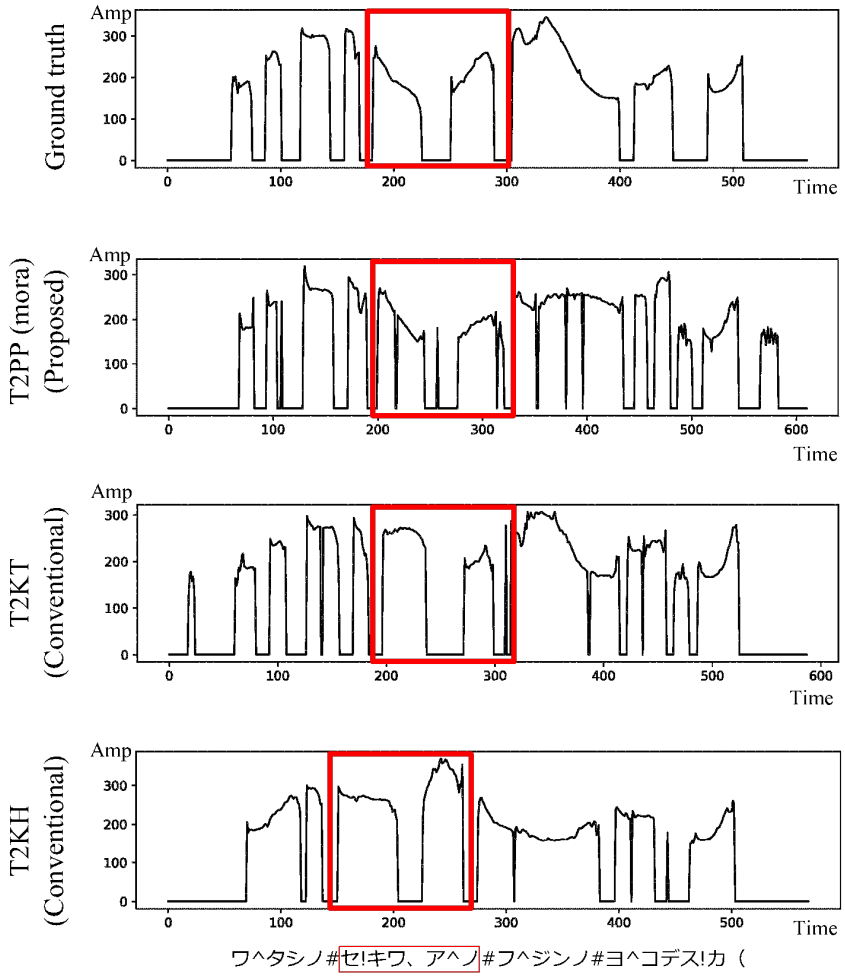


Figure 2.9: Comparison of F0 for “私の席は、あの婦人の横ですか。
(Is My Seat Next to That Lady?).”
（Copyright (C) 2021 IEICE, [5] Fig. 7）

Table 2.6: Comparison of F0 Correlation.

Systems	T2KH	T2KT	T2PP (phon.)	T2PP (mora)	ReWN
F0 correlation	0.23	0.27	0.40	0.38	0.48

Table 2.7: Similarity and Matching Rate of Strings.

	Strings' similarity	Whole strings' matching rate
PP (mora)	0.90	1.68%
KT	0.94	22.6%

and initial rising "^" symbols are effective. In the T2PP (mora) results of Figure 2.8, we can see a rapid falling of F0 corresponding to "セ!キワ" and a rapid rising of F0 corresponding to "ア^ノ." In T2PP (mora) results of Figure 2.8, we can see a reproduced silence corresponding to a pause "_". Overall, the results shown in these figures indicate that prosodic symbols can control acoustic features such as accent and pause information.

Table 2.6 compares the F0 correlations of T2PP (phon.), T2PP (mora), T2KT, and T2KH. The iterations and batch size of Tacotron 2 were 600,000 and 48, respectively. We used the WaveNet implementation [44] as the vocoder; The resynthesized results of the WaveNet vocoder are indicated as ReWN. The results of this experiment showed that the F0s of the two T2PPs were more similar than the F0s of the other systems. As well, T2PP (phon.) and T2PP (mora) with prosodic symbols had a higher evaluation value than T2KT. This result suggested that prosodic symbols replicate F0 and accentual features.

2.4.4. Comparing Manual and Automatic Generated Labels

Table 2.7 lists the matching rates of the manual and automatically generated labels. We prepared automatically generated and hand-edited PP labels made from 5,000 hand-edited FC labels [45]. They were made from 5,000 utterances of the JSUT corpus. We compared these labels and used the python difflib Sequence Matcher function [46] [47] to calculate the strings' similarity. Table 2.7 shows that the similarity in the case of using automatically generated PP labels is lower than that of kana. This experiment was conducted because it

```
Algorithm: Whole strings matching.
Input: S1, S2
Output: Truth value
function whole_strings_matching(S1,S2)
    count←0
    Iter←0
    for S1 ∈ l do
        if len(S1)=len(S2)
            if S1(Iter)=S2(Iter)
                count←count+1
            else
                break
            end if
        Iter←Iter+1
        else
            break
        end if
    end for
    if count=len(S1)
        return true
    end if
end function
```

Figure 2.10: Algorithm of Whole Strings' Matching.

is known that the accuracy of the labels affects the evaluation results [48].

The Python difflib function implements the Ratcliff-Obershelp algorithm [47]. The similarity of two strings S_1 and S_2 is determined by the formula:

$$D = \frac{2 \cdot \min\left(|S_1|, |S_2|\right)}{|S_1| + |S_2|} \qquad (1)$$

The whole strings' matching rate is calculated by counting the truth value of whole strings matching. Input two sets of sentences, S_1 and S_2, into the whole strings' matching function in Figure 2.10, counting the number of the truth value outputs generated by the function. Divide this number by the total number of sentences, and calculate the matching rate for the entire strings.

Table 2.8: Number of Miss-Conversions in Synthesized Speech.

	HAND	AUTO
Miss-synthesized	0 /50	5 / 50

Moreover, the matching rates were significantly different. The matching rate in the PP case was 1.68%, while that of kana was 22.6%. These results suggest that it is difficult to generate strings that perfectly match the hand-edited labels. Considering the difference in matching rates, predicting prosodic features is considered more difficult than kana. The prosodic features of Japanese depend on the context and potentially contain multiple patterns, so it is difficult to estimate them only from sentences without any acoustic features of speech.

2.4.5. Counting Errors in Miss-Synthesized Speech

Table 2.8 lists the number of synthesized utterances that were not generated correctly, comparing manually and automatically generated labels with synthesized speech. Seq2seq AMs have been reported to have synthesis errors that include deletions and repetitions of words [32]. These errors often occur at the end of a sentence. Even though a different sound is produced compared with the input sentence and it is a natural mistake for a sentence, it cannot be judged in subjective evaluations or from alignment errors. In this experiment, we prepared automatically generated PP labels (AUTO) and hand-corrected PP labels (HAND) made from the JSUT corpus 4,900 corrected FC labels [43]. The experiment was conducted using the ESPnet-Tacotron 2 [49] implementation and using the Griffin-Lim vocoder. There were 200 epochs, and we selected 50 texts and compared the input sentences with the synthesized speech by listening to them. We manually counted the errors in the synthesized speech. Table 2.8 shows that inputting the hand-corrected PP labels did not cause mis-synthesized speech. This means that PP labels are suitable input for seq2seq AM. Kanji conversion seemed to be the cause of the errors in the automatically generated PP labels; that is, PP labels did not cause the errors.

2.4.6. Subjective Evaluation

Four subjective evaluations were conducted using the 100 test samples as the evaluation stimuli. The evaluators were 200 speakers of standard Japanese (Tokyo dialect). The evaluated speech stimuli did not use any of the training data for the model. One audio sample was evaluated 20 times. Mean opinion scores (MOS) on a scale of 1-to-5 (1: bad, 5: excellent) and 95% confidence intervals were obtained from all the evaluators.

2.4.7. Effectiveness of Prosodic Symbols

The systems listed in Table 2.4 were compared in terms of the naturalness of the speech they produced. We estimated the mel-spectrograms (MELSPC) by using three seq2seq AM models and generated the audio by using the Griffin-Lim vocoder [50] (60 iterations). ReGL in the figure means re-synthesized audio created using the Griffin-Lim vocoder. The number of iterations and batch size of Tacotron 2 and Deep Voice 3 were 600,000 and 48, respectively. The number of iterations and batch size of Transformer-based TTS were 75,900 and 12, respectively. The results are summarized in Figure 2.11. All methods that inputted PP labels had evaluation values greater than those of KT and KH. It can be seen that the PP labels worked effectively on all of the seq2seq AM methods. These results suggest that prosodic symbols can be applied to various architectures with attention-based seq2seq AM.

In addition, we conducted an experiment with different types of phonetic symbols to evaluate the effectiveness of prosodic symbols in Figure 2.12. The systems listed in Table 2.4 and T2PP contained two types of phonetic symbol: Roman alphabet (phoneme) as T2PP (phon.) and katakana (mora) as T2PP (mora). The number of iterations and batch size of Tacotron 2 were 600,000 and 48. We generated the audio by using the WaveNet vocoder. The number of iterations and batch size for the vocoder were 200,000 and 48. The results are summarized in Figure 2.11. T2PP (phon.) and T2PP (mora) had evaluation values greater than T2KT and T2KH. It can be seen that the PP labels worked well for every type

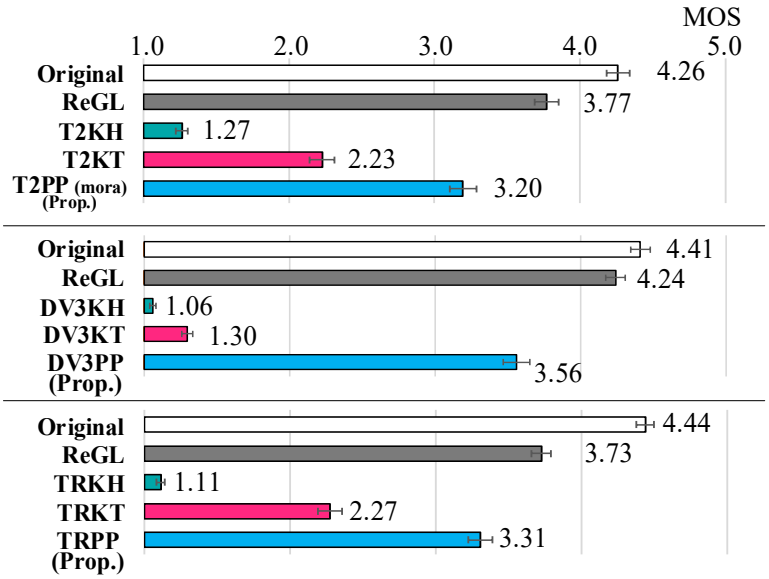


Figure 2.11: Effectiveness of Linguistic Phonological Symbols.
（Copyright (C) 2021 IEICE, [5] Fig. 8）

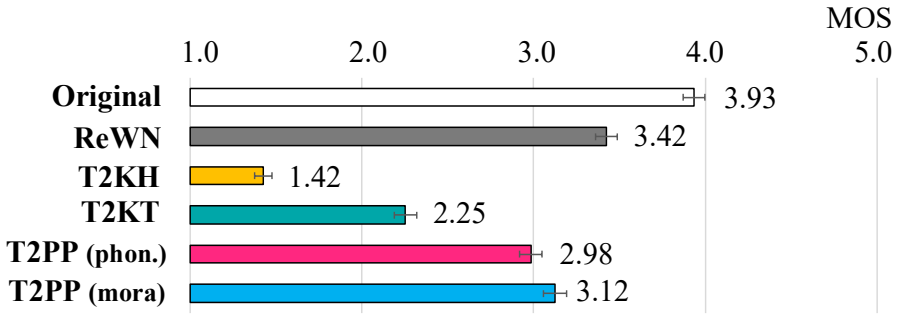


Figure 2.12: Effectiveness of Tacotron 2 and WaveNet with PP Labels.
（Copyright (C) 2021 IEICE, [5] Fig. 9）

of than T2KT and T2KH. It can be seen that the PP labels worked well for every type of phonetic symbol. In this experiment, the comparison between T2PP (phon.) and T2PP (mora) showed a significant difference in T2PP (mora). These results suggest that prosodic symbols can be applied to various phonetic symbols with seq2seq AM and that T2PP (mora) is more effective for naturalness than T2PP (phon.) in this method.

2.4.8. Comparison with Conventional SPSS

We built the TTS systems listed in Table 2.9. All of them used automatically generated

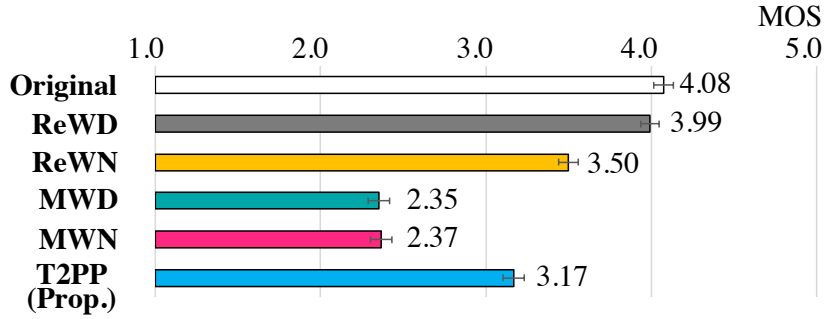


Figure 2.13: Comparison with Conventional SPSS.
(Copyright (C) 2021 IEICE, [5] Fig. 10)

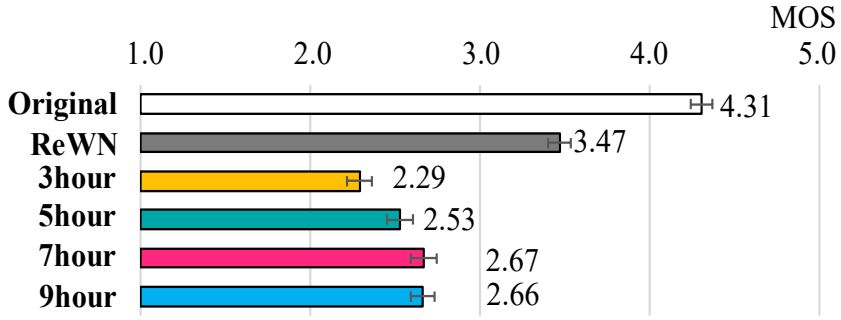


Figure 2.14: Effect of Changing the Volume of Training Data.
(Copyright (C) 2021 IEICE, [5] Fig. 11)

Table 2.9: Conventional Systems Used in the Experiments.

System	AM method	Acoustic feature	Waveform synthesis
ReWD	Re-synthesis	VOCODER	WORLD [27]
ReWN	Re-synthesis	MELSPC	WaveNet [44]
MWD	Merlin [51]	VOCODER	WORLD [27]
MWN	Merlin [51]	VOCODER	WaveNet [28]
T2PP	Tacotron 2 [43]	MELSPC	WaveNet [44]

labels as training data. The conventional SPSS [51] employed Open JTalk [40] and Julius [52] [52] as a front-end. There were 281 errors due to forced alignment by Julius. Consequently, the training set had 7,324 sentences instead of 7,596, the test set had 91 instead of 100, and the evaluation object had 30. We used WORLD [28] as the vocoder, 60-dimensional mel-cepstral coefficients (MCCs), 2-dimensional band periodicities (BAPS), log F0 at 5 msec frame intervals (the acoustic features of VOCODER), and three recurrent hidden layers; each hidden layer had 512 LSTM (long short-term memory) units

as duration and acoustic models.

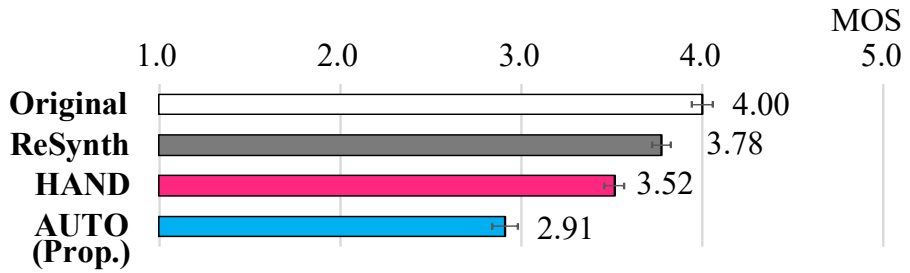


Figure 2.15: Effectiveness of Auto-Generated Labels.
（Copyright (C) 2021 IEICE, [5] Fig. 12）

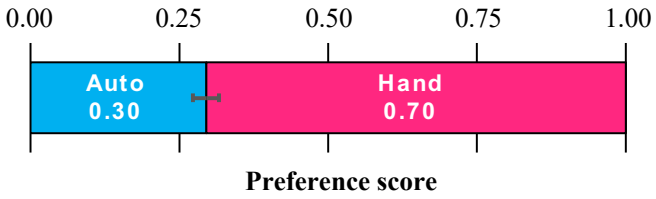


Figure 2.16: Results of Pairwise Comparison of Auto-Generated Labels and Hand-Edited Labels with 95% Confidence Interval.
（Copyright (C) 2021 IEICE, [5] Fig. 13）

The results are summarized in Figure 2.13. Our method scored significantly higher than MWN and MWD, which are conventional SPSS methods. This result confirms the effectiveness of the proposed method without time alignment information. The reason for ReWN's score being lower than that of ReWD is that synthesized speech rarely contains unnatural sounds. A similar issue has been reported in experiments with neural vocoders.

2.4.9. Effect of Changing the Volume of Data

We subjectively evaluated the effect of changing the volume of the training data of T2PP (mora). We generated the audio by using the WaveNet vocoder. The number of iterations and the batch size in this experiment were 200,000 and 48. Figure 2.14 shows that the evaluation values gradually increased with the number of hours of training. The systems trained on sets with less data tended to yield sentences that were perceived as mis-phrased and poorly accented. As the learning data was increased, the reproducibility of the phonetic

and prosodic features increased, as did their naturalness.

2.4.10. Comparison of Using Automatically Generated Labels and Hand-Corrected Labels

We subjectively evaluated the effectiveness of automatically generated PP labels in Figures 2.15 and 2.16. For this experiment, we prepared automatically generated (AUTO) and hand corrected (HAND) labels made from 5,000 hand-corrected FC labels [30]. We generated the audio from the predicted mel-spectrograms by using the Griffin-Lim vocoder. It is logical to think that this experiment should have no effect on the evaluation result for any type of vocoder, because it does not affect high/low pitch accent information. For this reason, we used the Griffin-Lim vocoder. The corpus contains 5,000 utterances of the JSUT corpus; the training set includes 4,900 utterances, the test set 100. We randomly selected 30 sentences. The evaluators were 60 speakers of standard Japanese (Tokyo dialect). The evaluated speech stimuli did not use any of the training data for the model. One audio sample was evaluated 30 times. There were 200 epochs, and we conducted a 1-to-5 evaluation (MOS score) and a pairwise comparison of the naturalness of the synthesized speech.

Figures 2.14 and 2.15 show that hand corrected PP labels was rated higher than automatically generated ones. These results suggest that the correctness of the labels affects the results of the evaluation. As described in Section 2.4.4, automatic labels contain errors due to miss-conversions, while prosodic features, which contain accent information, are difficult to estimate; these issues caused the low evaluations. Mistakes in the prosodic features might have affected the reproducibility of the accents, etc., and reduced their naturalness.

2.5. Discussion

2.5.1. Experimental Findings

We confirmed the effectiveness of automatic generated labels for pitch-accent language in Japanese. The experiments showed that symbols can be used to control the accentual

acoustic feature. Figures 2.8 and 2.9 confirm the controllability of the prosodic feature by comparing mel-spectrograms and F0. Until now, there has been no generic accent control method for Japanese seq2seq AM. The proposed method potentially solves the accentual control problems. Moreover, the prosodic features work with multiple seq2seq methods; this is in contrast to conventional SPSS, which requires extensive manual labor for correcting pronunciations by using labels describing the boundaries of phonemes. Although, the automatically generated labels were evaluated as worse than the hand-corrected ones, this could be attributed to miss-converted labels. Because the overall similarity is high, the phoneme alignments were correct, but the reproduced accents contained errors.

In Section 2.4.5, it was shown that PP labels are suitable for the input of seq2seq AMs. Seq2seq AMs rarely cause miss-syntheses that are different from the sentence input at the end of the word. These sounds are often linguistically correct, but they do not express the same sound as the input text [32]. We experimentally confirmed how often errors occurred through the contribution of hand-edited labels [45] and found that no such errors occurred in the case of PP labels in the experiment described in Section 2.4.5. Prosodic features affect tone and pitch, and in Section 2.4.6, we showed it is possible for these acoustic features to be accurately represented by putting them between phonemes. For this reason, it may be possible that problems that cannot be learned with phonemes only, such as accurate accents, can be learned accurately by using PP labels.

2.5.2. How to Improve the Automatically Generated Labels

The production cost of controllable TTS is high in the conventional SPSS method using manual labeling. As described in Section 2.4.6, we conducted subjective evaluation experiments confirming that seq2seq AM with our method produces more natural speech than SPSS does. However, while this would solve the cost problem, the quality of the automatically generated labels remained lower than that of the hand-corrected labels. If this problem in pitch-accent languages can be solved, high-quality TTS can be realized without the need to use hand-edited labels.

The results of the subjective evaluation suggested that the prosodic symbols are effective. In Section 2.4.4, despite that automatically generated label contained miss-

converted symbols, the PP labels estimated tones of accent, etc. The experimental results showed that the prosodic symbols had other functions besides controlling phonemes. They also can be used to control the prosodic features, which do not correspond to the phonemes. Moreover, as shown in Section 2.4.6, the prosodic features could be adapted to seq2seq AMs that have not been proposed yet, and this might mean that we do not need to develop any new seq2seq AM models to input accent information. Accordingly, we can avoid developing new structures for accent control to support new seq2seq AMs.

In Section 2.4.6, it was shown that the prosodic features can control the accent, but we could not estimate the labels correctly. It was found that the Tacotron 2 used in this experiment could produce relatively correct speech even though there were some miss-converted labels. Moreover, it was found that an increase in accent estimation accuracy improved the evaluation results. Increasing the accuracy of accent estimation will lead to higher quality in the future.

The prosodic features potentially contain multiple expressions, so finding methods of linguistic analysis for accurately determining phonemes and acoustic analysis for accurately estimating accent remains an issue. In this study, we used linguistic analysis only for automatically generating labels; in the future, we would like to confirm the effectiveness of adding a prosodic symbol estimation method using acoustic analysis. In summary, the results of this study suggest that the prosodic features are effective for Japanese seq2seq AM. Moreover, because the input of seq2seq AM is merely symbols, the prosodic features may work in languages other than Japanese.

2.5.3. Points to be Confirmed in Future Experiments

The evaluation of ReWN described in Section 2.4.7 gave an unclear result. In this regard, we used mel-spectrograms extracted from raw audio to train the ReWN model; nevertheless, the result was not good. The reproduced consonants in ReWN may not have been good. In particular, confusion in the perception of the reproduced consonants may have influenced the results of the evaluation. However, ReWN did not seem to affect the results of the proposed method, because ReWN seemed to have no effect on the experimental results of T2PP (mora), MWD or MWN.

2.6. Conclusion

We proposed a method to control prosodic features by inserting symbols representing these features between phonetic symbols in various architectures with attention-based seq2seq AM. The addition of prosodic symbols resulted in more accurate replication of accents, pauses, and sentence ending acoustic expressions and improved the evaluation value compared with methods inputting only plain phoneme sequences. Moreover, we found that the proposed method in combination with a front-end could automatically generate speech without imposing a large annotation workload. But an evaluation of the automatically generated labels and hand-edited labels showed that the hand-edited labels were still better, so the automatic estimation method for labels should be improved. The naturalness of the speech indicated in the evaluations was higher than that of conventional SPSS. Our method has the potential for application to various languages and to work in as yet undeveloped seq2seq acoustic modeling methods.

Chapter 3

Phonemes and Prosodic Feature Estimation

3.1. Introduction

We present a method for estimating Japanese prosodic features [3]. Our proposed approach accurately discerns not just accent features, but also phonemes, pauses, and accented phrases. It serves a dual purpose: automatic label generation for speech synthesis [3] and aiding language learners in speech production training [53] [54]. The former serves as an automatic label generation method for speech synthesis [3], while the latter aims to assist in training for pitch accent languages [38]. In speech training, this enables learners to visualize prosodic information in their speech and utilize it for practice. While prior pitch-based accent recognition methods faced challenges in symbol consistency, our speech synthesis model exhibits promise for prosodic symbol training. To gather ample labeled training data, we propose a self-supervised acoustic model integrating the Transformer [55]. This model achieves precise recognition of Japanese readings and accents with just around 5 hours of training data, achieving an impressive 96% recognition rate, showcasing its precision.

3.2. Conventional Method and Its Problems

We utilize two conventional approaches: the initial being an acoustic modeling and PP label conversion method [3], while the latter involves the seq2seq acoustic modeling approach

[56]. Traditional Japanese accent identification typically relies on analyzing F0 [10]. However, Japanese accent rules are rooted in linguistic principles, particularly high/low pitch accents, rather than technical specifications. Hence, the F0 of speech data may not inherently signify high/low pitch accents. Although there exists a method for analyzing F0 and estimating accents through machine learning [10], the recognition rate for accents remains modest, approximately 76%. In the domain of speech synthesis, the challenge is compounded by the utilization of PP labels in the training data, resulting in an even lower recognition rate due to the necessity to encompass both accent patterns and the reading of kana characters. The recognition rate of the method outlined in [10] is inferior to that observed in the preliminary experiment.

Various alternative conventional methods have been proposed. Ishi introduced an analysis method which organized the classification tree of F0 [58], yet evaluation outcomes are subpar. Hatano proposed a method for DNN-based pitch accent estimation from F0 trajectories [7], however, evaluation outcomes are also unsatisfactory. Koriyama proposed the TTS method employing semi-supervised prosody modeling with a deep gaussian process model [59] [60]. Nevertheless, this method appears proficient in recognizing accent features but falters in estimating that information due to TTS. Yufune proposed a method based on variational autoencoder (VAE) and vector quantization (VQ)-VAE-based method, in which linguistic and acoustic features are inputted into the encoder [61]. It can merely recognize accents on a word-by-word basis, and its applicability is constrained because it necessitates the addition of linguistic information. Our investigation illustrates that prosodic symbols can effectively regulate high/low pitch accents in speech synthesis (Chapter 2). Common to these conventional methods is their inability to estimate prosodic features solely from speech; they solely estimate accentual features. All methods necessitate supplementary linguistic information, which encompasses speech. Additionally, all methods can only recognize accurate features. The proposed approach aims to redress these shortcomings.

Hence, we employ two conventional methods which exclusively train speech data for evaluating effectiveness. The first traditional method (Figure 3.1) entails the cascaded acoustic modeling to estimate sentences containing a mixture of kanji and kana characters [32] and the speech synthesis language processing unit to estimate PP labels (Chapter 2).

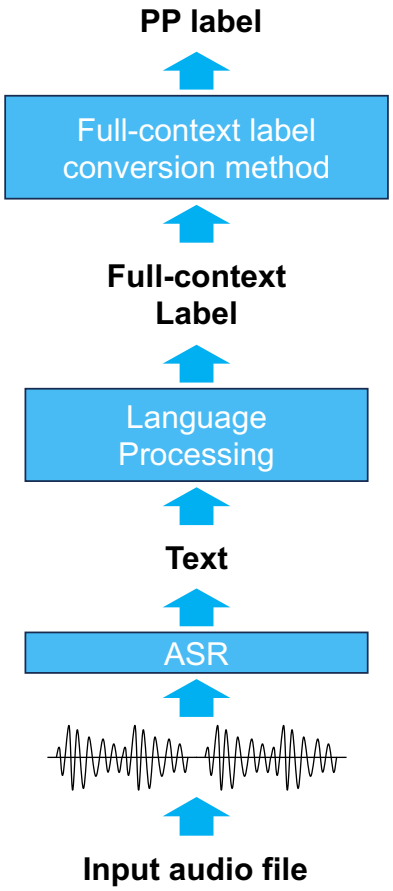


Figure 3.1: The First Conventional Method of ASR and Full-Context Label Conversion Method.

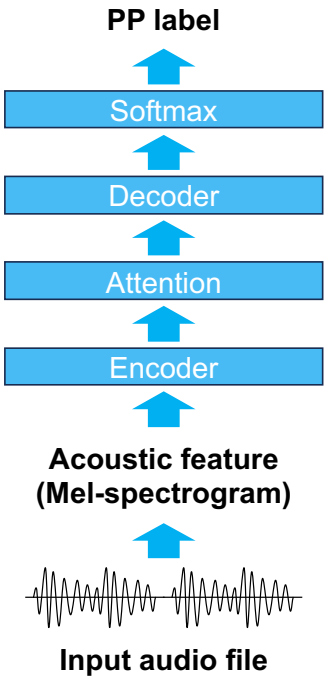


Figure 3.2: The Second Conventional Method of Seq2seq Acoustic Modeling Based-PP Label Estimation.

Furthermore, the first method, which is incapable of directly estimating actual accents from speech data, also yields a diminished recognition rate. The second traditional method (Figure 3.2) involves training and estimating PP labels using seq2seq acoustic modeling, typically utilized as a speech recognition method [60] [61] [62].

3.3. Proposed Method

The study herein focuses on phonemes and prosodic feature recognition, employing self-supervised acoustic modeling [60]. Additionally, practical illustrations of TTS applications are introduced. To validate the efficacy of Chapter 2, we propose that seq2seq acoustic

modeling can estimate PP labels from speech. Nonetheless, it's crucial to acknowledge that seq2seq acoustic modeling typically demands several thousand hours of training data [56], significantly surpassing the usual 10 hours requisite for speech synthesis (Section 2.4.8). Consequently, while models trained with this method can identify PP labels, the recognition rate is anticipated to be comparatively low. We showcase the effectiveness of the proposed method in contrast to these two approaches, emphasizing its effectiveness even with a restricted amount of data.

3.3.1. Phonetic and Prosodic Feature Estimation

We presented an approach for integrating PP labels into Japanese speech synthesis in Chapter 2. Waveform and PP labels constitute language-specific data pairs, often challenging to collect in abundance for speech recognition purposes. Hence, we directed our attention towards Baevski's self-supervised learning acoustic modeling method [36] [37] [38] [61], adept at learning from scanty data using self-supervised learning techniques [61]. Moreover, recognition outcomes frequently exhibit phoneme inaccuracies. In our endeavor to rectify character strings plagued by phoneme errors, we endeavored to enhance recognition precision by employing Text-to-Text Transformer (T5) [62].

Experimentation revealed that this methodology could discern pitch accent through its application in PPF estimation. Typically, the training data available for TTS purposes barely exceeds ten hours, insufficient for the extensive data demands of PPF estimation, which necessitates several hundred hours. Nonetheless, our self-supervised learning approach facilitated the development of a recognition method with just five hours of data in our trials. To bolster recognition accuracy with minimal data, we leveraged a publicly available pre-trained model of self-supervised learning acoustic modeling [63], trained on a vast corpus of 56,000 hours of speech data encompassing 53 languages [64].

Wav2vec 2.0 stands as a self-supervised acoustic modeling speech recognition paradigm, harnessing 56,000 hours of extensive speech data spanning 53 languages, trained sans labels, serving as a pre-training model. Employing CNNs, Wav2vec 2.0 trains on speech data to derive latent speech representations $\boldsymbol{Z}$ (Figure 3.3). Renowned for its adeptness in speech recognition with few-shot inputs, Wav2vec 2.0 boasts high-quality

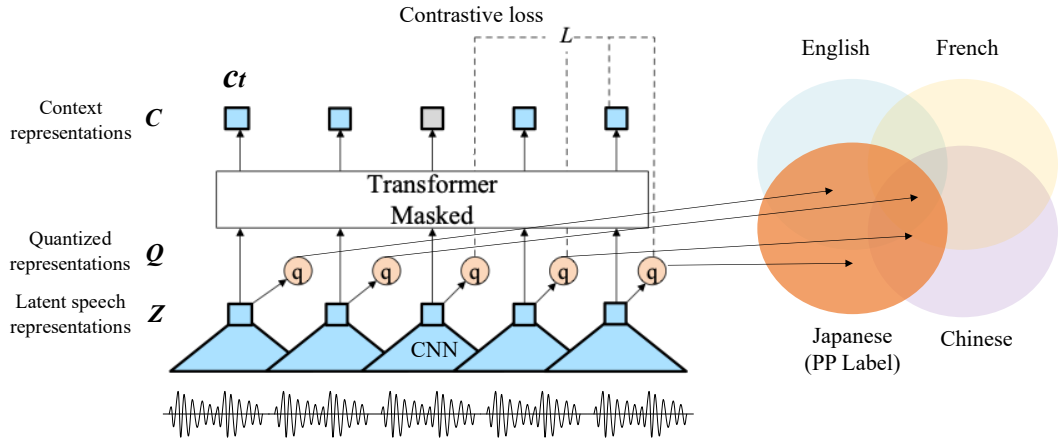


Figure 3.3: Fine-tuning Quantized Representation on Information of Phonemes from Various Languages with Pre-training.

outcomes. The Transformer component within the model autonomously refines contextual representations $\boldsymbol{C}$ by inferring from one of the $\boldsymbol{Z}$, following the training paradigm akin to BERT [61], a notable self-supervised learning technique in natural language processing (NLP).

In our proposed methodology, PP labels, scarce in availability, undergo fine-tuning with a pre-training model to serve as a PP labels recognizer. Through fine-tuning, the quantized representation $\boldsymbol{Q}$ assimilates phoneme data from diverse languages accrued during pre-training, culminating in superior recognition accuracy despite limited speech data.

Despite achieving commendable recognition accuracy, this method still contends with prevalent errors in recognizing consonants in Japanese speech. These consonantal recognition anomalies are frequently assuaged by the language processing unit in speech recognition systems. We posit that implementing a mechanism to diminish errors in consonantal phonemes could further enhance recognition rates. To this end, we advocate for a system wherein the Transformer rectifies the most probable phoneme errors through training on recognition outcomes inclusive of phoneme errors from authentic speech inputs to the recognizer, complemented by manually corrected labels (Figure 3.4). Augmenting the T5 training data involved devising a phonetic and accentual error generator via random character deletions and consonant substitutions.

```
Algorithm: Consonants error generator.
Input: Consonants set (Cp ), replacement_probability
Output: Random replaced consonants set (Cr )
function cons_random(Cp )
    for i∈ range(len(Cp )) do
        if random.random() < replacement_probability
            Cr ←random.choice(Cp )
        else
            Cr .remove
        end if
    end for
    return Cr
end function
```

Figure 3.4: Algorithm of Consonant Error Generator.

However, it's important to acknowledge that Transformers typically necessitate millions of sentences for training. In our scenario, we have access to only approximately 20,000 sentences, equivalent to about 10 hours of data. To mitigate this constraint, we have chosen to augment our sentence data utilizing data augmentation as a pre-training mechanism. Specifically, we randomly extract symbols from Japanese texts, substitute consonants with alternative consonants, and/or remove phonemes themselves. This strategy employs data augmentation to amplify the training dataset. The subsequent equations and algorithm demonstrate data augmentation for substituting consonants in PP labels to train the pre-trained model. Consonantal phonemes (***Cp***), randomly selected consonants (***Cr***), vowels (***V***), and prosodic features (***P***) constitute string sets:

$$\boldsymbol{Cp} = \{cp_1, cp_2, \cdots, cp_l\} = \{/k/, /s/, /t/, \cdots, /sh/, /\emptyset/\}$$
$$\boldsymbol{V} = \{v_1, v_2, \cdots, v_m\} = \{/a/, /i/, /u/, /e/, /o/\}$$
$$\boldsymbol{P} = \{p_1, p_2, \cdots, p_k\} = \{/\mathit{Initial\ rising}/, /EOS/, /\mathit{Pause}/,$$
$$/\mathit{Accent\ phrase\ boundary}/, /\mathit{Accent\ nucleus}/, /\emptyset/\}$$

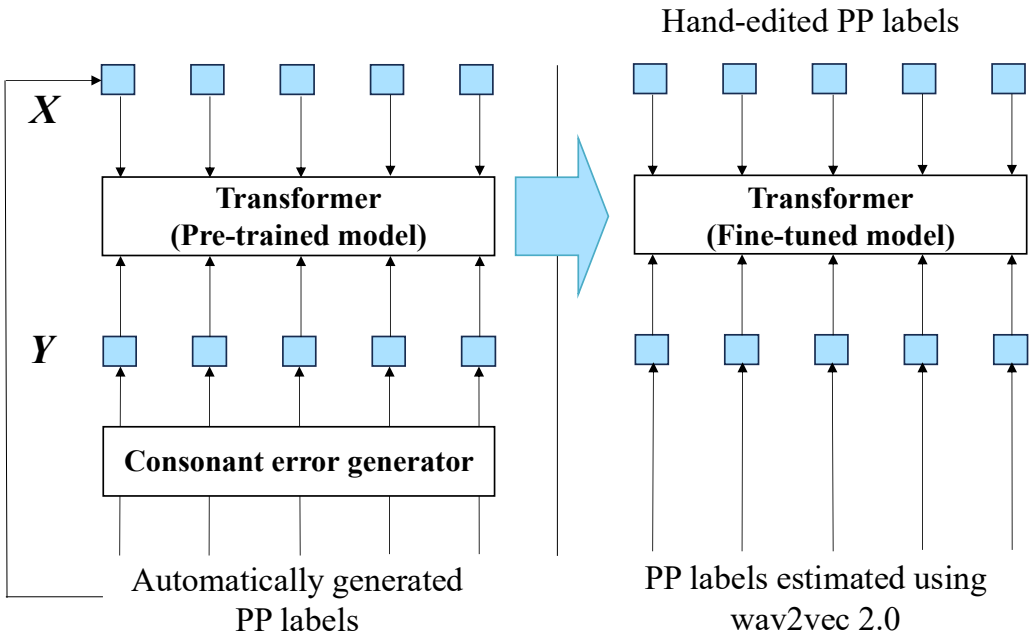


Figure 3.5: Phoneme-Error-Correction Transformer.

The algorithm delineating the consonantal error generator from ***Cp*** to ***Cr*** is depicted in Figure 3.4. The input sequence of tokens $\boldsymbol{X}$ $(x_1, x_2, \cdots, x_N)$ and a target sequence $\boldsymbol{Y}$ $(y_1, y_2, \cdots, y_N)$ are illustrated below:

$$\boldsymbol{X} = \begin{cases} \in \boldsymbol{Cp} & i = 1,4,7,\cdots,N-2 \\ \in \boldsymbol{V} & i = 2,5,8,\cdots,N-1 \\ \in \boldsymbol{P} & i = 3,6,9,\cdots,N \end{cases} \quad (1)$$

$$\boldsymbol{Y} = \begin{cases} \in \boldsymbol{Cr} & i = 1,4,7,\cdots,N-2 \\ \in \boldsymbol{V} & i = 2,5,8,\cdots,N-1 \\ \in \boldsymbol{P} & i = 3,6,9,\cdots,N \end{cases} \quad (2)$$

The condition probability of the pre-trained model for $\boldsymbol{X}$ and $\boldsymbol{Y}$ are shown below:

$$P(\boldsymbol{X}|\boldsymbol{Y}) = \prod_{n=1}^{N} P(y_n|(y_1, y_2, \cdots, y_{n-1}), \boldsymbol{X}) \quad (3)$$

In Figure 3.5, we undertook fine-tuning utilizing observational data on the pre-trained model. We trained the PP labels inferred from the waveform with the SSL-based AM alongside the manually edited PP labels. This enables SSL to refine observed phonetic discrepancies from the fine-tuned wav2vec 2.0 estimation model.

Subsequently, the pre-training model is trained with this augmented data, and fine-tuning ensues using phoneme error data generated by the actual recognizer to formulate a

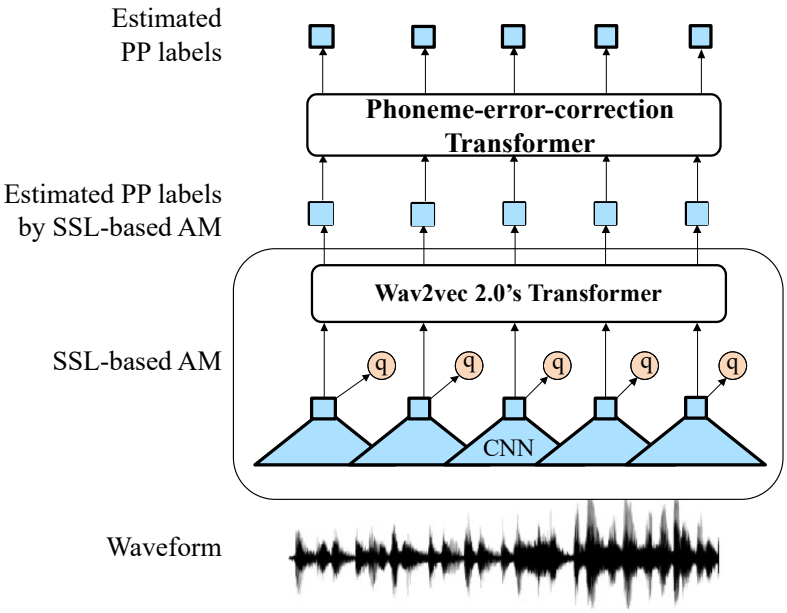


Figure 3.6: SSL-Based Phonetic and Prosodic Labels Estimation Method.

phoneme error correction Transformer. Therefore, we propose a combined approach involving self-supervised learning acoustic modeling and a phoneme error correction Transformer to attain high-quality PP label recognition (Figure 3.6).

3.4. Evaluation Experiment

To evaluate the effectiveness of the proposed method, we performed an evaluation experiment on the accuracy of recognizing PPF. We used the recorded speech of in-house anchors in a studio booth and manually corrected PPF in speech datasets for fine-tuning the pre-trained model of wav2vec 2.0.

3.4.1. Experimental Conditions for Phonetic and Prosodic Feature Recognition

We prepared a dataset of four males and a dataset of 3 females. The manuscript of dataset consisted of news, weather forecast, and lifestyle information. The sampling frequency was 16 kHz, and the bit rate was 16 bits. In addition, the text data for pre-training of T5 was used for automatic generated PPF, which our proposed G2P (Section 2.3.4) [5] generated from 631,014 sentences of news script obtained from NHK NEWS WEB [65] in the period

from April 2018 to April 2021. For a wav2vec 2.0 pre-trained model, we used XLSR-Wav2Vec2 [63] which contains approximately 56,000 hours of speech data for 53 languages as data for pre-training. We performed fine-tuning against this pre-trained model using speech and manually corrected Japanese kana characters and prosodic symbols and then performed model training. Additionally, we trained the T5 used for correcting PPF errors by automatically creating characters for the 631,014 sentences of news manuscripts using our proposed G2P tool (Section 2.3.4). The following data augmentation of T5 processing was performed against the above data to create training data.

- PPF was deleted at a rate of 5% or less.
- Consonants of PPF were substituted at a rate of 10% or less

We fine-tuned the pre-trained model generated with data used for pre-training as described above by using a training set consisting of 23,024 sentences of manually corrected PP label. In conformance with the properties of PP labels, prosodic symbols consisted of initial rising, accent nucleus, accentual phrase boundary, pause, and end of sentence, which related accentual and pause information. PPF estimation by wav2vec 2.0 was taken to be proposed Prop. 1 and correcting the PPF estimated by wav2vec 2.0 by T5 for correcting phoneme errors was taken to be proposed Prop. 2. Additionally, since the amount of training data in the dataset for TTS was insufficient to use the speech-recognition method for comparison purposes, we decided for this experiment to convert speech into PPF by using a pre-trained model for Japanese speech recognition released by ESPnet ASR [5] that uses seq2seq speech-recognition method. We also used the process of automatically converting that speech into PPF using our proposed G2P [5] as a conventional method (Conv.) for comparison.

Table 3.1: Comparison of Proposed Method and Seq2seq Acoustic Modeling-Based PP Label Estimation.

	Method	CER %
Conv.	Seq2seq AM + Open JTalk	22.6
Prop. 1	Wav2vec 2.0	8.5
Prop. 2	Wav2vec 2.0 + Transformer (T5)	4.7

3.4.2. Experiment 1

We used the speech in the dataset of one male and that of one female (2541 sentences, 5.69 hours) to fine-tune wav2vec 2.0 in proposed methods Prop. 1 and Prop. 2. We also used character strings of manually corrected PP labels (23,024 sentences) to fine-tune T5 for correcting phoneme errors in proposed method Prop. 2. As for the test dataset, we used the dataset of two males and one female (1558 sentences, 3.73 hours). In the experiment, we calculated the character error rate (CER) between the estimated labels obtained by each method and ground truth and compared results. We use seq2seq acoustic modeling-based PP label recognition method as conventional method, referred to Section 3.3 and Figure 3.2.

Experimental results are listed in Table 3.1. Proposed methods Prop. 1 and Prop. 2 had lower values of CER than the conventional method. Moreover, comparing proposed methods Prop. 1 and Prop. 2, the value for proposed method Prop. 2 that incorporated T5 for correcting phoneme errors was lower, which demonstrated the effectiveness of using T5.

3.4.3. Experiment 2

In this experiment, we examined the impact of varying amounts of training data, ranging from 1.0 to 20.0 hours. We employed a corpus containing both male and female speech for fine-tuning in wav2vec 2.0, without the use of a Transformer for error correction. The experimental conditions were as follows: M001 and F001 served as variable training sets. M002, F002, and M004 (consisting of 1558 sentences totaling 3.73 hours) were utilized as

Table 3.2: Evaluation Results of Increased Data Volume.

Hour	CER %
1.0	9.3
2.5	8.3
5.0	7.5
10.0	7.7
20.0	7.8

the test set. For the phoneme error correction Transformer, 631,014 sentences from news manuscripts sourced from NHK NEWS WEB were employed to simulate phoneme errors by randomly deleting letters and replacing consonants. The data enhanced pretrain model was then applied to 23,024 manually corrected sentences for label fine-tuning on the training set. Self-supervised acoustic modeling was pre-trained using Facebook's “XLSR-Wav2Vec2 [63].”

The evaluation results, presented in Table 3.2, illustrate the outcomes with different data quantities. According to the experimental findings, the highest performance was achieved with 5 hours of data.

3.4.4. Experiment 3

In this experiment, we conducted a comparison between the proposed method and seq2seq acoustic modeling by utilizing seq2seq acoustic modeling to train the PP labels. This confirmed the effectiveness of self-supervised learning. Since seq2seq typically requires large amount of training data, we prepared 10 hours of data, surpassing the volume used in the experimental results obtained in Chapter 2, in order to match the experimental conditions. While the conventional method of seq2seq acoustic modeling was trained solely on this dataset, the proposed method made use of a publicly available pre-training model with 10 hours of fine-tuning data. The training data for both methods is same. We use seq2seq acoustic modeling-based PP label recognition method as conventional method,

Table 3.3: Comparison of Proposed Method and Seq2seq Acoustic Modeling-Based PP Label Estimation.

	Conventional: Seq2seq acoustic modeling-based PP label recognition method	Proposed: Self-supervised acoustic modeling-based PP labels recognition method
CER %	10.6	6.8

referred to Section 3.3 and Figure 3.2.

The experimental conditions were as follows: F001 (1834 sentences, 3.00 hours), M001 (1834 sentences, 4.8 hours), M002 (55 sentences, 0.11 hours), M003 (114 sentences, 0.22 hours), and M006 (663 sentences, 2.01 hours), resulting in a total of 10.14 hours of training data. For the test set, we used M004, M005, F002, and F003 (total 250 sentences, 0.53 hours). Self-supervised acoustic modeling was pre-trained using Facebook's XLSR-Wav2Vec2 [63].

The experimental results revealed that the proposed method achieved a significantly lower CER compared to the conventional method (Table 3.3). This confirms that the self-supervised learning method enhances the recognition rate, even when the data volume is limited.

3.5. Discussion

We demonstrated the effectiveness of the proposed method in Experiment 1. Through this experiment, we gained knowledge on how prosodic symbols could be estimated by using “Wav2vec 2.0 + Transformer (T5)” with high accuracy from only speech. Conventional methods cannot estimate PPF that reflects acoustic features. We found that the proposed method Prop. 2 could estimate high-quality PP labels. According to Experiment 2, the evaluation value of the training data is constant at over 5 hours. We proved that a small amount of training data is sufficient to obtain good label estimation results.

3.6. Conclusions

We have introduced a method capable of estimating PP labels from speech. Our approach includes PP labels estimation method that leverages a large-scale pre-training model and a self-supervised learning based acoustic modeling. We have confirmed the effectiveness of this proposed method in accurately estimating labels through self-supervised learning acoustic modeling and a phoneme error correction Transformer. In 2023, there has been no method available for directly estimating PPF from speech. This method presents applications for labels estimation in the proposed Japanese TTS method and for the training of Japanese speech.

Chapter 4

Speech Synthesis with Adjustable Acoustic Features

4.1. Introduction

Deep learning-based text-to-speech is used in various situations and the sound quality is close to that of humans. We previously develop DNN-based SPSS [1] method and also developed our DNN-TTS for controlling speaking style, speaker, speech rate, pitch, and intonation. More specifically, this method enables the changing of specific speaking styles, such as news speaking style that mimics various style. In this chapter, we propose the method of controlling speaking style, speaker, speech rate, pitch, and intonation, and evaluate its effectiveness.

4.2. Conventional Method and Its Problems

In 2013, the DNN-TTS method based on SPSS [1] was proposed and showed significant improvement in speech quality. This conventional method could only train one style and speaker; therefore, it could not train a single model even if there was a corpus of various speakers and styles. In 2018, Hojo proposed a speech synthesis method that can switch speakers [11]. However, a method for controlling speaking style has not been proposed. Additionally, it lacked the ability to adjust speech rate, pitch, and intonation. The conventional method could only process speech in the waveform after speech synthesis had taken place, making it impossible to adjust these features at the intermediate processing

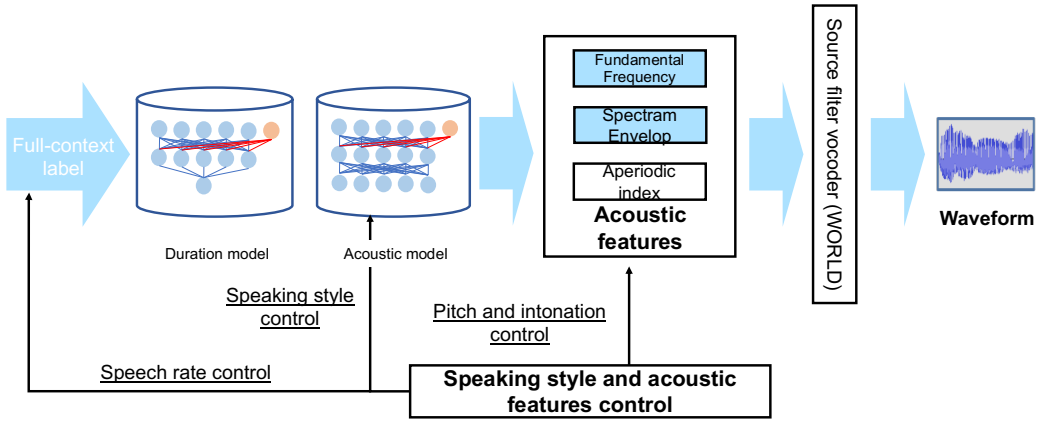


Figure 4.1: Overall Speaking Style and Acoustic Features Control in DNN-Based Statistical Speech Synthesis.

acoustic feature stage before input to the vocoder [28].

4.3. Proposed Method

The proposed method combines the controllable speaking style method with the control of speech rate, pitch, and intonation. An overview of the proposed method is presented in Figure 4.1. The speech rate [66] is regulated by modifying the full-context label [24]. Pitch and intonation are manipulated by directly editing the acoustic features generated by the acoustic model [28]. The adjusted acoustic features are then fed into the source filter vocoder to generate speech. As speaking style cannot be quantitatively adjusted, the style is managed by modifying the DNN model.

4.3.1. DNN Speech Synthesis with Controllable Speaking Style

We propose the method for controlling speaking styles which developed for our DNN-TTS method and has reproduced speaking styles and speaker identity. We develop a method to control each of the style and speaker code by inputting them into both the duration model and the acoustic model (Figure 4.2). Typically, only one speaker and style can be trained for each feature. However, by controlling and training with these features, it is possible to train a single model for speech with a variety of features. This approach increases the amount of speech data that can be learned by a single model and is expected to improve

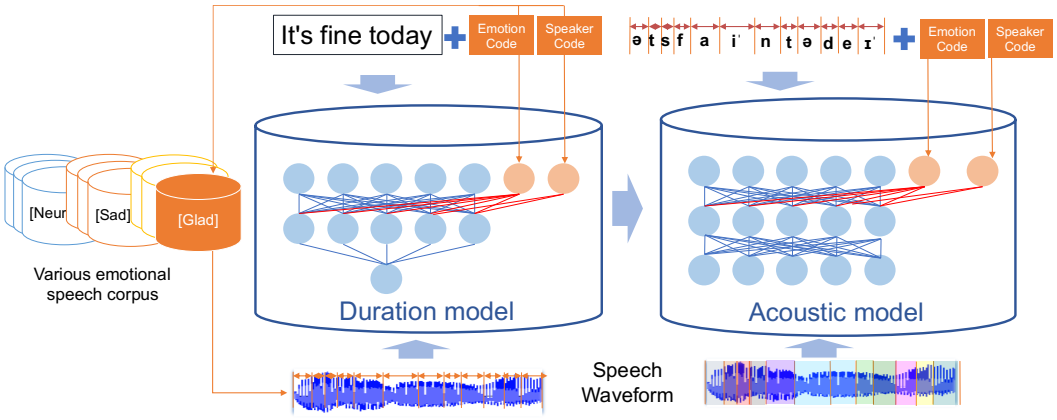


Figure 4.2: Controlling Speaking Styles for DNN-Based Statistical Speech Synthesis.

overall naturalness. When controlling emotion and speaker variables, incorporate the emotion code and speaker code into both the duration model and acoustic model. Subsequently, through training a corpus divided by emotion and speaker, it becomes possible to learn how to adjust both emotion and speaker.

4.3.2. Method for Controlling Speech Rate, Pitch, and Intonation

We propose a DNN-TTS method utilizes a source-filter vocoder [28] in the waveform synthesis part. This vocoder conducts numerical simulations of speech, combining sound sources such as the vocal cords and vocal tract [28]. By adjusting the parameters of the acoustic features, proposed method is possible to modify speech rate, pitch, and intonation. These adjustments are controlled by changing the parameters of the DNN models employed in the DNN speech synthesis processing. Figure 4.3 illustrates our method of controlling acoustic feature parameters. Using this approach, acoustic features parameters are modified to directly reflect the acoustic features estimated from the DNN models. This allows for the alteration of speaking styles, minimizing sound quality deterioration before the vocoder step that generates the waveform. The advantages of these methods empower us to produce various types of high-quality speech. The acoustic features of the source filter vocoder,

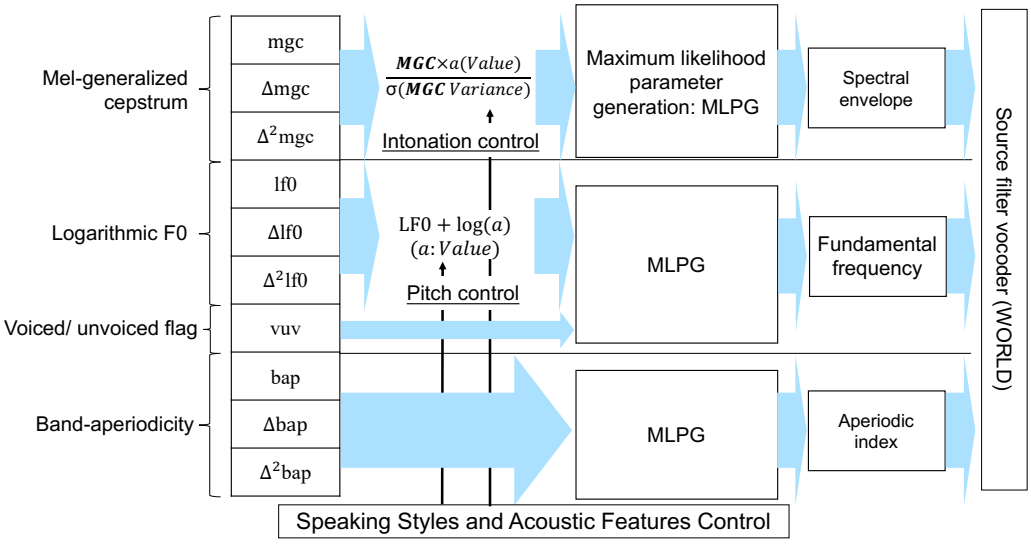


Figure 4.3: Procedure for Speaking Style and Acoustic Features Control.

generated by the acoustic model, are decomposed into mel-generalized cepstrum (MGC) [67], logarithmic F0 (LF0) [68], voice/unvoiced (VUV) flag, and band-aperiodicity (BAP) [67]. The acoustic features of the filter vocoder are further decomposed into MGC and LF0. The acoustic features, adjusted by the equation shown in Figure 4.3, can be calculated using maximum likelihood parameter generation (MLPG) [69] and then converted to a spectral envelope, F0 and aperiodic index [28]. Each feature is then input into the source filter vocoder to generate speech.

4.4. Evaluation Experiments

We conducted two experiments to evaluate the effectiveness of the proposed method: the first experiment aimed to examine the method's effectiveness in training multiple corpora into a single model, and the second experiment evaluated the effectiveness of the proposed method in a specific task. We assessed whether the speech generated by the proposed method improves the understanding of Japanese by non-native learners. Through these two experiments, we demonstrated the effectiveness of the proposed method's model and provided examples of its successful application.

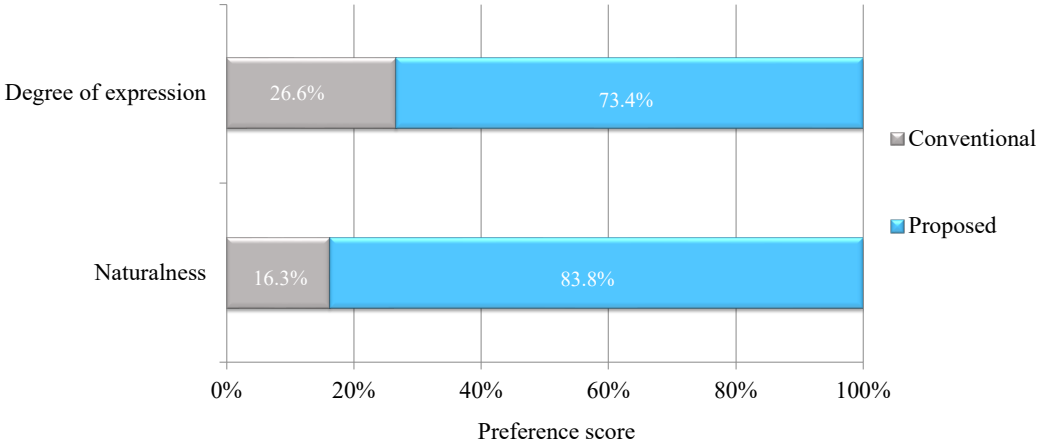


Figure 4.4: Comparison of DNN-Based Statistical Speech Synthesis Trained with Multiple Styles Using the Paired Comparison Method.

In the first experiment, subjective evaluation experiment on the effectiveness of speech various style control methods. We examine whether the quality increases with the total amount of data, even when dealing with a corpus of limited size. This experiment aims to test the effectiveness of using speaker codes for training, especially with an increase in training data. The pairwise comparison method was employed to assess speech generated by training on multiple corpora versus speech generated by a model trained on a single style and speaker corpus. The parameters under scrutiny include the “degree of expression” and “naturalness.” The training corpus consists of male speech, 5 speakers, and 4 emotions (normal, happy, angry, sad), with 495 sentences in each corpus. Four subjects, randomly presented with headphones, participated in the evaluation, which involved 20 sentences distinct from the training corpus and employed two methods of synthesized sounds.

The results are summarized in Figure 4.4. Regarding the “degree of expression,” reflecting emotional expression in speech, the proposed method scored 73.4%, whereas the conventional method scored 26.6%. The proposed method demonstrated a significantly higher evaluation in terms of the emotional aspect of the speech. Furthermore, for “naturalness,” assessing sound quality and other aspects of speech, the proposed method achieved a significantly higher score (83.3%) compared to the conventional method (16.3%). These findings indicate that the proposed method received significant effectiveness in terms of emotional expression and overall naturalness. The model, which

trained multiple styles using style and speaker codes, was highly evaluated.

In the second experiment, synthesized speech with manuscripts expressed in easy Japanese, which were checked and corrected by an editor and released on News Web Easy, were evaluated in terms of preference. We used an internal corpus of 21 hours of speech (11,716 sentences) uttered by a professional female narrator. The corpus was split into 11,612 sentences for training and 104 samples for testing. They conducted subjective evaluations of understanding on a crowd sourcing website. The evaluated speech stimuli did not use any of the training data for the models. One audio sample was evaluated 20 times. We prepared five types of speech, which were recorded as RAW and DNN TTS synthesized files. RAW is a recording of a speech spoken slowly by a news anchor. TTS_Direct is just TTS synthesized speech. ORG is speech published daily on the homepage and composed by Japanese language teachers. These teachers manipulate the speech rate by adjusting it for slower speech, shifting the higher pitch of the speech, and manually correcting the accent. SS is a speech that converts TTS_Direct by using a generic speech rate conversion application [66]. The speech length of ORG and SS was set by the utterance length of RAW and was 25% longer than the average utterance length of TTS_Direct. PH is a speech that has a higher pitch than the average pitch frequency of TTS_Direct. The pitch shift rate of PH was determined by ORG's pitch shift rate. The test set included 125 samples for evaluation and five samples for training the participants as

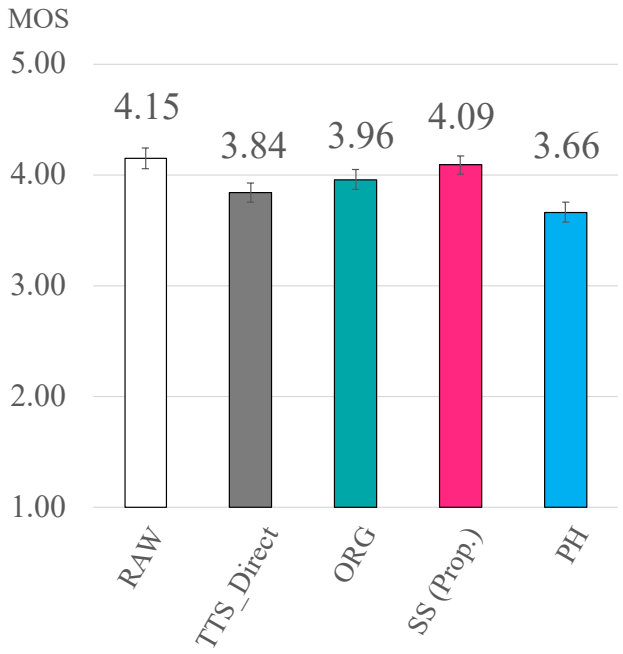


Figure 4.5: Synthesized Speech with Manuscripts Expressed in Easy Japanese News.

evaluators. The mean opinion scores (MOSs) on a scale of 1 to 5 (1: bad, 5: excellent) and 95% confidence intervals were obtained from all participants.

The results are summarized in Figure 4.5. The ratings of ORG and SS (Prop.) tended to be higher than that of TTS_ Direct. There was a significant difference between TTS_Direct and SS (Prop.). These results indicate that the effectiveness of slow speech rate is high. And there was no significant difference between ORG and SS (Prop.). This suggests that the difference between speech rate conversion using a vocoder and generic speech rate conversion methods did not affect the experimental results. These results indicate that the understanding of a manuscript increases as the speech rate decreases, and there is little effect of pitch shift. In this section, we demonstrate the effectiveness of adjusting acoustic features in specific tasks. Speech synthesis finds diverse applications, and its efficacy cannot be solely assessed by evaluation methods focused on speech naturalness. Here, we investigate the method's effectiveness in a task designed for learners of Japanese—one of the more constrained applications of speech synthesis. Our findings reveal that, in this particular task, the proposed method enhances Japanese comprehension

by slowing down speech rate in the acoustic features. This underscores a correlation between speech rate and Japanese comprehension.

4.5. Conclusion

In this chapter, we have developed DNN-based SPSS with speaking style and acoustic features control. Through this method, we demonstrate that a diverse corpus consisting of various speakers and speaking styles can be effectively trained within a single model. Furthermore, by directly manipulating the acoustic features of the DNN-TTS, it is possible to make adjustments with minimal degradation in sound quality. In the specific scenario, the evaluation significantly improved compared to the evaluation conducted before adjusting the acoustic features.

Chapter 5

Conclusion

In this thesis, we present three achievements. The first pertains to Japanese speech synthesis. We have demonstrated that prosodic symbols can effectively control the acoustic features of pitch accents in the seq2seq acoustic modeling. It is known that seq2seq-based TTS can be achieved by training phonemes in various languages, and it was successful in training hiragana and katakana in Japanese. However, Japanese speech synthesis faced challenges due to the inability to reproduce pitch accents. The proposed method was found to effectively reproduce pitch accents, thereby enabling Japanese speech synthesis in seq2seq-based TTS. As the input features consist of symbol sequences, the proposed method exhibits a high affinity with other NLP methods [70]. Experiments have shown the effectiveness of this method.

The second achievement is the successful utilization of PP labels estimation method from speech. Prior to this research, there was a notable absence of a method to estimate phonetic and prosodic features from speech using the self-supervised learning based-acoustic model estimation approach. The amount of data for PP labels is small. Therefore, a method had to be devised to achieve inference even with a limited dataset, utilizing self-supervised learning.

The third achievement revolves around speech synthesis with adjustable acoustic features. We have developed speech synthesis with speaking style control, enhancing the quality of synthesized speech. This quality improvement is a direct result of training a single model with a diverse corpus encompassing various speaking styles. Furthermore, by manipulating the acoustic features, which represent intermediate features in DNN-TTS

systems, our research has achieved the ability to adjust the features of speech. These methods have also been seamlessly integrated into the seq2seq framework. This method has also been applied to seq2seq based TTS.

Utilizing the findings from this three research, NHK has implemented Japanese speech synthesis in its news programs [71]. Seq2seq-based TTS can accurately reproduce the speech of professional news anchors. The anchors employing this text-to-speech method feature prominently in NHK's daily nationwide news programs, reaching a substantial audience and thereby establishing them as highly influential.

Furthermore, our demonstration establishes the applicability of this method to a diverse range of seq2seq-based TTS models through training on symbol sequences. Given its versatility and high-quality performance, the method has been incorporated into major open-source projects for speech signal processing, including ESPnet [32]. Consequently, it has found widespread adoption in variety of research and development initiatives as a Japanese seq2seq-based TTS, currently recognized as the mainstream approach for Japanese speech synthesis. Presently, speech-based large language models (LLMs) are experiencing a surge in development, with numerous emerging speech synthesis methods. However, the majority of LLMs are grounded in the Transformer architecture, making this method highly applicable. Consequently, the implementation of LLS based-multilingual text-to-speech synthesis [72] [73] also integrates Japanese grapheme-to-phoneme (G2P) method [5] [40], encompassing our proposed ideas. Given the intricacies of the Japanese language, characterized by multiple readings of kanji characters and linguistic properties that influence phonetic and prosodic features in speech symbols, our thesis methods hold promise for future applications.

Bibliography

[1] H. Zen, A. Senior, and M. Schuster, “Statistical Parametric Speech Synthesis Using Deep Neural Networks,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] K. Kurihara and M. Abe, “Automatic Commentary Training Text to Speech System for Broadcasting,” in *International Broadcasting Convention (IBC) Technical Papers*, 2022.

[4] K. Kurihara, N. Seiyama, T. Kumano, T. Fukaya, K. Saito, and S. Suzuki, “‘AI News Anchor’ with Deep Learning-Based Speech Synthesis,” *SMPTE Motion Imaging Journal*, vol. 130, no. 3, pp. 19–27, 2021.

[5] K. Kurihara, N. Seiyama, and T. Kumano, “Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS,” *IEICE Transactions on Information and Systems*, vol. E104-D, no. 2, pp. 302–311, 2021.

[6] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2Wav: End-to-End Speech Synthesis,” in *International Conference on Learning Representations (ICLR)*, 2017.

[7] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, “Investigation of Enhanced Tacotron Text-to-Speech Synthesis Systems with Self-Attention for Pitch Accent Language,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6905–6909.

[8] Y. Sagisaka and H. Sato, “Accentuation Rules for Japanese Word Concatenation,”

IEICE Transactions on Information and Systems, vol. 66, no. 7, pp. 849–856, 1983.

[9] NHK (Japan Broadcasting Corporation), "*NHK's New Dictionary of Japanese Pronunciation and Accentuation,*" *NHK Broadcasting Culture Research Institute*, 2016. (in Japanese)

[10] H. Hatano, A. Albin, R. Wang, and C. Ishi, "Classification of Japanese Accent Types Using Machine Learning: A Comparison of Native Speakers' and Learners' Utterances from Word List and Read-Aloud Tasks," in *General Meeting of the Phonetic Society of Japan*, 2018, pp. 49–53. (in Japanese)

[11] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-Based Speech Synthesis Using Speaker Codes," *IEICE Transactions on Information and Systems*, vol. E101-D, no. 2, pp. 462–472, 2018.

[12] K. Kurihara, N. Seiyama, T. Kumano, T. Fukaya, K. Saito, and S. Suzuki, "'AI News Anchor' with Deep Learning-Based Speech Synthesis," *SMPTE Motion Imaging Journal*, vol. 130, no. 3, pp. 19–27, 2021.

[13] K. Tamaoka, S. Makioka, S. Sanders, and R. Verdonschot, "Www.kanjidatabase.com: A New Interactive Online Database for Psychological and Linguistic Research on Japanese Kanji and Their Compound Words," *Psychological Research*, vol. 81, 2017.

[14] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[15] H. Zen and A. Senior, "Deep Mixture Density Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3844–3848.

[16] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *International Conference on Machine Learning (ICML)*, 2018, pp. 7471–7480.

[17] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. Le, "Tacotron: Towards End-to-End Speech Synthesis," in *International Speech Communication Association (INTERSPEECH)*, 2017, pp. 4006–4010.

[18] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop," *arXiv preprint arXiv:1707.06588*, 2017.

[19] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling Text-To-Speech with Convolutional Sequence Learning," in *International Conference on Learning Representations (ICLR)*, 2018.

[20] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural Speech Synthesis with Transformer Network," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 6706–6713.

[21] T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Impacts of Input Linguistic Feature Representation on Japanese End-to-End Speech Synthesis," in *ISCA Speech Synthesis Workshop (SSW)*, 2019, pp. 166–171.

[22] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial Investigation of an Encoder-Decoder End-to-End TTS Framework Using Marginalization of Monotonic Hard Latent Alignment," *arXiv preprint arXiv:1908.11535*, 2019.

[23] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-Time Neural Text-to-Speech with Sequence-to-Sequence Acoustic Model and Waveglow or Single Gaussian WaveRnn Vocoders," in *International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1308–1312.

[24] Nagoya Institute of Technology, "An Example of Context Dependent Label Format for Hmm-Based Speech Synthesis in Japanese," HMM/DNN-based Speech Synthesis System (HTS), [Online]. Available: http://hts.sp.nitech.ac.jp/archives/2.3/HTS-demo_NIT-ATR503-M001.tar.bz2.

[25] S. Shechtman and A. Sorin, "Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities," in *ISCA Speech Synthesis Workshop (SSW)*,

2019, pp. 275–280.

[26] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, “Transformer Based Grapheme-to-Phoneme Conversion,” in *International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2095–2099.

[27] Y. Yu, X. Si, C. Hu, and J. Zhang, “A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures,” *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[28] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications,” *IEICE Transactions on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[29] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-Dependent Wavenet Vocoder,” in *International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1118–1122.

[30] A. van den 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.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 1-9.

[32] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, “ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-To-Speech Toolkit,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7654–7658.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1-9.

[34] D. P. Kingma and P. Dhariwal, “Glow: Generative Flow with Invertible 1x1 Convolutions,” in *Advances in Neural Information Processing Systems*, 2018, pp. 10235–10244.

[35] J.J. Venditti, “Japanese Tobi Labelling Guidelines,” *Ohio State University Working Papers in Linguistics*, pp. 123-162, 1997.

[36] J. E. Shoup, “Phonological Aspects of Speech Recognition,” *Trends in Speech Recognition*, pp. 125–138, 1980.

[37] H. Fujisaki and K. Hirose, “Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese,” *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[38] N. Minematsu, S. Kobayashi, S. Shimizu, and K. Hirose, “Applying Conditional Random Fields to Japanese Morphological Analysis,” in *International Speech Communication Association (INTERSPEECH)*, 2012, pp. 2561–2564.

[39] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2004, pp. 230–237.

[40] Nagoya Institute of Technology, “Open JTalk,” [Online]. Available: http://open-jtalk.sourceforge.net/.

[41] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT Corpus: Free Large-Scale Japanese Speech Corpus for End-To-End Speech Synthesis,” *arXiv preprint arXiv:1711.00354*, 2017.

[42] R. Yamamoto, “Deepvoice3_pytorch,” [Online]. Available: https://github.com/r9y9/deepvoice3_pytorch.

[43] R. Mamah, “Deepmind’s Tacotron 2 Tensorflow Implementation,” [Online]. Available: https://github.com/Rayhane-mamah/Tacotron-2.

[44] R. Yamamoto, “WaveNet_Vocoder,” [Online]. Available: https://github.com/r9y9/wavenet_vocoder.

[45] K. Tomoki, “Jsut-Label,” [Online]. Available: https://github.com/sarulab-speech/jsut-label.

[46] The Python Software Foundation, “Difflib — Helpers for Computing Deltas,” [Online]. Available: https://docs.python.org/3/library/difflib.html.

[47] J. W. Ratcliff and D. Metzener, “Pattern Matching: The Gestalt Approach,” *Dr. Dobb's Journal*, vol. 13, no. 7, p. 46, 1988.

[48] H. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, “Investigating Accuracy of Pitch-accent Annotations in Neural Network-based Speech Synthesis and Denoising Effects,” in *International Speech Communication Association (INTERSPEECH)*, 2018, pp. 37–41.

[49] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, and N. Chen, “ESPnet: End-To-End Speech Processing Toolkit,” in *International Speech Communication Association (INTERSPEECH)*, 2018, pp. 2207-2211.

[50] D. Griffin and L. Jae, “Signal Estimation from Modified Short-Time Fourier Transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[51] Z. Wu, O. Watts, and S. King, “Merlin: An Open Source Neural Network Speech Synthesis System,” in *ISCA Speech Synthesis Workshop (SSW)*, 2016, pp. 202–207.

[52] A. Lee, T. Kawahara, and K. Shikano, “Julius - An Open Source Real-Time Large Vocabulary Recognition Engine,” [Online]. Available: https://github.com/julius-speech/segmentation-kit.

[53] B. Muradás-Taylor, “Accuracy and Stability in English Speakers’ Production of Japanese Pitch Accent,” *Language and Speech*, vol. 65, no. 2, pp. 377–403, 2022.

[54] X. Li, C. T. Ishi, and R. Hayashi, “Prosodic and Voice Quality Feature of Japanese Speech Conveying Attitudes: Mandarin Chinese Learners and Japanese Native,” in *Speech Prosody*, 2020, pp. 41–45.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All You Need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 1–15.

[56] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1-9.

[57] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, and Y. Wu, “Conformer: Convolution-Augmented Transformer for Speech Recognition,” *arXiv preprint arXiv:2005.08100*, 2020.

[58] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/Attention Architecture for End-to-End Speech Recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[59] A. Graves, N. Jaitly, and A. Mohamed, “Hybrid speech recognition with deep bidirectional LSTM,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 273–278.

[60] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[62] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the Limits of Transfer Learning with A Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[63] Q. Xu, A. Baevski, and M. Auli, “Simple and Effective Zero-Shot Cross-Lingual Phoneme Recognition,” *arXiv preprint arXiv:2109.11680*, 2021.

[64] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised Cross-lingual Representation Learning for Speech Recognition,” in *International Speech Communication Association (INTERSPEECH)*, 2020.

[65] NHK, “NHK News Web,” [Online]. Available: https://www3.nhk.or.jp/news. (in Japanese)

[66] A. Imai, N. Tazawa, T. Takagi, T. Tanaka, and T. Ifukube, “A New Touchscreen Application to Retrieve Speech Information Efficiently,” *IEEE Transactions on Consumer Electronics*, vol. 59, no. 1, pp. 200–206, 2013.

[67] S. Furui, “Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[68] K. Yu and S. Young, “Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[69] H. Zen and H. Sak, “Unidirectional Long Short-Term Memory Recurrent Neural Network with Recurrent Output Layer for Low-Latency Speech Synthesis,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4470–4474.

[70] N. Kakegawa, S. Hara, M. Abe, and Y. Ijima, “Phonetic and Prosodic Information Estimation from Texts for Genuine Japanese End-To-End Text-to-Speech,” in *International Speech Communication Association (INTERSPEECH)*, 2021, pp. 3606–3610.

[71] K. Kurihara, “Research and Practical Application of AI Anchors Using Japanese Text-to-Speech Method,” *The Journal of the Institute of Image Information and Television Engineers*, vol. 78, no. 2, 2024. (in Japanese)

[72] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, and J. Li, “Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling,” *arXiv preprint arXiv:2303.03926*, 2023.

[73] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, and J. Li, “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers,” *arXiv preprint arXiv:2301.02111*, 2023.

Appendix

Publications

Publications Related to the Thesis

Journal Papers

[P1] Kiyoshi KURIHARA, Nobumasa SEIYAMA and Tadashi KUMANO, "Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS," *IEICE Transactions on Information and Systems*, vol. E104-D, no. 2, pp. 302-311, 2021, doi: 10.1587/transinf.2020EDP7104. → **Chapter 2.**

[P2] Kiyoshi KURIHARA, Nobumasa SEIYAMA, Tadashi KUMANO, Takashi FUKAYA, Kazunari SAITO and Satoshi SUZUKI, ""AI News Anchor" with Deep Learning-Based Speech Synthesis," *SMPTE Motion Imaging Journal* vol. 130, no. 3, pp. 19-27, 2021, doi: 10.5594/JMI.2021.3057703. → **Chapter 4.**

International Conference Paper (Peer-Reviewed)

[P3] Kiyoshi KURIHARA and Mayumi ABE, "Automatic Commentary Training Text to Speech System for Broadcasting," in *International Broadcasting Convention (IBC) Technical Papers*, Paper Session: How AI is Advancing Media Production, 12 pages, 2022. → **Chapter 3.**

Publications Non-Related to the Thesis

Journal Paper

[P4] Kiyoshi KURIHARA, Atsushi IMAI, Nobumasa SEIYAMA, Toshihiro SHIMIZU, Shoei SATO, Ichiro YAMADA, Tadashi KUMANO, Reiko TAKO, Taro MIYAZAKI, Manon ICHIKI, Tohru TAKAGI and Hideki SUMIYOSHI, "Automatic Generation of Audio Descriptions for Sports Programs," *SMPTE Motion Imaging Journal*, vol. 128, no. 1, pp. 41-47, 2019, doi: 10.5594/JMI.2018.2879261.

[P5] Kiyoshi KURIHARA, "Research and Practical Application of AI Anchors Using Japanese Text-to-Speech Method," *The Journal of the Institute of Image Information and Television Engineers*, vol. 78, no. 2, 2024. (in Japanese, to be published)

International Conference Papers (Peer-Reviewed)

[P6] Nobuaki MINEMATSU, Fuki YOSHIZAWA, Tadashi KUMANO, Kiyoshi KURIHARA and Daisuke SAITO, "Comparison of Accent Theories of Japanese Using E2E Speech Synthesis in Terms of Their Effectiveness for Learners to Acquire Natural Prosody," in *International Symposium on Applied Phonetics*, pp. 53-58, 2021, doi: 10.21437/ISAPh.2021-9.

[P7] Kiyoshi KURIHARA, Nobumasa SEIYAMA, Tadashi KUMANO, Takashi FUKAYA, Kazunari SAITO and Satoshi SUZUKI, ""AI News Anchor" with Deep Learning-Based Speech Synthesis," in *SMPTE Annual Technical Conference and Exhibition*, pp. 19-27, 2020, doi: 10.5594/M001915.

[P8] Tadashi KUMANO, Tohru TAKAGI, Manon ICHIKI, Kiyoshi KURIHARA, Hiroyuki KANEKO, Tomoyasu KOMORI, Toshihiro SHIMIZU, Nobumasa SEIYAMA, Atsushi IMAI and Hideki SUMIYOSHI, "Generation of Automated Sports Commentary from Live Sports Data," in *The IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1-4, 2019, doi: 10.1109/BMSB47279.2019.8971879.

[P9] Manon ICHIKI, Toshihiro SHIMIZU, Atsushi IMAI, Tohru TAKAGI, Mamoru IWABUCHI, Kiyoshi KURIHARA, Taro MIYAZAKI, Tadashi KUMANO, Hiroyuki KANEKO, Shoei SATO, Nobumasa SEIYAMA, Yuko YAMANOUCHI and Hideki SUMIYOSHI, "Study on Automated Audio Descriptions Overlapping Live Television Commentary," in *International Conference on Computers Helping People*, pp. 220-224, 2018, doi: 10.1007/978-3-319-94277-3_36.

[P10] Kiyoshi KURIHARA, Atsushi IMAI, Hideki SUMIYOSHI, Yuko YAMANOUCHI, Nobumasa SEIYAMA, Toshihiro SHIMIZU, Shoei SATO, Ichiro YAMADA, Tadashi KUMANO, Reiko TAKO, Taro MIYAZAKI, Manon ICHIKI, Tohru TAKAGI, Susumu OSHIMA and Koji NISHIDA, "Automatic Generation of Audio Descriptions for Sports Programs," in *International Broadcasting Convention (IBC) Technical Papers*, Paper Session: AI is Here and the Machines are Hungry to Learn, 8 pages, 2017.

Endnote

This thesis is based on "Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS" [P1], by the same author, which appeared in the Proceedings of "Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS," Copyright (C) 2021 IEICE. The material in this thesis was presented in part at the proceedings of "Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS," [P1] and all the figures of this paper are reused from [P1] under the permission of the IEICE.