Rakugo Speech Synthesis: Toward Speech Synthesis That Entertains Audiences

by

Shuhei KATO

Dissertation

submitted to the Department of Informatics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The Graduate University for Advanced Studies, SOKENDAI

March 2021

Committee

Advisor	Dr. Junichi YAMAGISHI
	Professor of the National Institute of Informatics, Tokyo, Japan
Subadvisor	Dr. Isao ECHIZEN
	Professor of the National Institute of Informatics, Tokyo, Japan
Subadvisor	Dr. Imari SATO
	Professor of the National Institute of Informatics, Tokyo, Japan
Examiner	Dr. Takao KOBAYASHI
	Professor Emeritus of the Tokyo Institute of Technology, Tokyo, Japan
Examiner	Dr. Hiroki MORI
	Associate Professor of the Utsunomiya University, Tochigi, Japan

Abstract

Conventional speech synthesis research has focused on transferring information which the speech should have, such as content and speakers' emotions, personality, intention, accurately to listeners. Setting this purpose is reasonable considering that speech is a kind of media. Today, some speech synthesis systems can successfully produce speech as natural as human speech, albeit in the case of using well-articulated read speech.

However, the role of speech is not just information transfer. For example, verbal entertainment, including rakugo, on which we focus in this thesis, entertains audiences through the medium of speech. In other words, speech has a role of stirring listeners' emotion. This role has not been focused on enough in speech synthesis research, but we believe it is a good time to attempt to realize speech synthesis that entertains audiences because some modern speech synthesis systems have an ability to produce speech as natural as human speech, as mentioned above, albeit in the case of read-aloud speech.

In this thesis, we attempt to build rakugo speech synthesis as a challenging example of speech synthesis that entertains audiences. Rakugo is a traditional Japanese form of verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling. Although rakugo has a more than 300-year history, it is popular even today in Japan. In rakugo, a performer plays multiple characters, and conversations or dialogues between the characters make the story progress.

First, we built a large rakugo speech database for our study because there were no rakugo speech databases usable to train speech synthesis models. Most commercial rakugo recordings, thousands of which we can easily access, are live recordings that include noise and reverberation, whereas even modern speech synthesis cannot yet properly model such noisy and reverberant speech; therefore, we needed to build

a rakugo speech database. We recorded performances by a shin-uchi (first-rank professional) performer to train speech synthesis models, and performances by professional performers at various levels including the shin-uchi performer to evaluate synthesized speech. We not only transcribed the pronunciation of the recorded speech but also appended context labels to each sentence for better modeling of the speech.

Using the database, we modeled rakugo speech using segment-to-segment neural transduction (SSNT) based speech synthesis. The SSNT-based model has no soft attention network. An attention network maps the encoder and decoder time steps in a sequence-to-sequence speech synthesis model. Sequence-to-sequence models greatly improve the quality of speech synthesis, but attention networks occasionally cause unacceptable errors during synthesis. Since rakugo speech is far more diverse and casually-pronounced than speech ordinarily used for building speech synthesis, an attention network may cause errors more frequently; therefore using SSNT-based speech model, which has no attention networks, will be reasonable for modeling rakugo speech. We also used global style tokens (GSTs), which is a style transfer mechanism for sequence-to-sequence models, or manually labeled context features to enrich speaking styles of synthesized rakugo speech. Although the combination of the SSNT-based model and GSTs produced somewhat natural, character-distinguishable, and content-understandable speech, the mean opinion scores for this speech were just around 3 through a listening test.

For further improvement, we attempted Tacotron 2, a state-of-the-art speech synthesis model, and an enhanced version of it with self-attention to better consider long-term dependency. We also used GSTs, manually labeled context features, or the combination of them. Through a listening test, we found that state-of-the-art TTS models could not yet reach the professional level, and there were statistically significant differences in terms of naturalness, distinguishability of characters, understandability of the content, and even the degree of entertainment; nevertheless, the results of the listening test provided some interesting insights: 1) we should not focus only on naturalness of synthesized speech but also the distinguishability of characters and the understandability of the content to further entertain listeners; 2) the f_o expressivity of synthesized speech is poorer than that of human speech, and more entertaining speech should have richer f_o expression.

Lastly, we proposed a novel methodology for evaluating rakugo speech and

conducted a listening test to investigate how the level of rakugo speech synthesis compares to professional rakugo performers at various levels. Through a listening test, we found that the level of speech synthesis did not reach that of human professionals. On the other hand, the results suggested that we also should at least improve the f_o expression of speech synthesis to catch up with human professionals.

Although there is room for improvement, we believe this thesis is an important stepping stone toward achieving entertaining speech synthesis at the professional level.

Acknowledgements

I would like to express my gratitude to:

- Prof. Junichi YAMAGISHI, Ph.D., who is my supervisor and advisor. Since I decided to take an entrance examination to The Graduate University for Advanced Studies, SOKENDAI, he has made great efforts to support and guide me. I am very honored to be his student;

- Prof. Isao ECHIZEN, Ph.D., and Prof. Imari SATO, Ph.D., who are my subadvisors and members of the dissertation committee. They discussed my work with me excitedly and gave me a lot of significant advice;

- Prof. Takao KOBAYASHI, Ph.D., and Assoc. Prof. Hiroki MORI, Ph.D., who are members of the dissertation committee. They also discussed my work with me passionately. Their precise comments led my work to success;

- Mr. Yusuke YASUDA, who is a student at SOKENDAI, Kanagawa, Japan, and a member of Yamagishi Laboratory at the National Institute of Informatics (NII), Tokyo, Japan. His fine implementation of speech synthesis systems and a lot of technical advice strongly helped me to progress my work;

- Dr. Xin WANG, who is a Project Researcher at NII and also a member of the Yamagishi Lab. His clear comments sharpened my work;

- Dr. Erica COOPER, who is also a Project Researcher at NII and also a member of the Yamagishi Lab. She helped me as a professional researcher of speech synthesis and a beginner to Japanese culture;

- Dr. Shinji TAKAKI, who was a Project Researcher at NII, a member of the Yamagishi Lab., and is currently a Project Researcher at the Nagoya Institute of Technology, Nagoya, Japan. I could not start my work without his kind support and advice;

- Ms. Mika SASAKI and Ms. Makiko KUWAHARA, who is or used to be an Assistant Administrative Staff at the Yamagishi Lab. Their fine administrative work allowed me to focus on research;

- Master Yanagiya Sanza, who is a professional rakugo performer. He cooperated with us to record his professional and excellent rakugo performance. He also helped me to understand rakugo itself more accurately;

- Ryutei Ichido-san and (former) Yangiya Kogoto-san, who is or used to be a professional rakugo performer. They also cooperated with us to record their professional and exciting rakugo performance;

- Dr. Gustav Eje HENTER, Dr. Jaime LOLENZO-TRUEBA, Dr. Hieu-Thi LUONG, Dr. Yi ZHAO, and Mr. Haoyu LI, who is or used to be a member of the Yamagishi Lab. I really enjoyed the days spent with them;

- Dr. Fuming FANG, Mr. Hong Huy NGUYEN, and Ms. Ngoc-Dung Thi TIEU, who are members of Echizen Laboratory, NII, a collaborative laboratory with the Yamagishi Lab. Regular group meetings with them deepened my understanding of statistical learning;

- Prof. Masaru KITSUREGAWA, Ph.D., who is the Director General of NII. His interest, support, and advice to our rakugo speech synthesis project encouraged and helped me well;

- Dr. Deoksu NA, who was my colleague at Voiceware, Seoul, Korea (currently ReadSpeaker Korea; a company of HOYA Group, Tokyo, Japan). When I worked for HOYA Service Corporation, Tokyo, Japan, as a speech synthesis engineer, he kindly passed on knowledge and skills about processing speech to build a speech database for speech synthesis at the commercial level. I could not build a rakugo speech database for my work without these knowledge and skills;

- Ms. Maiko ONOE, who was my colleague at HOYA Service Corp. and HOYA Corporation, Tokyo, Japan. She was the most reliable colleague as an annotator and evaluator of speech. The outstanding quality of her work I saw was a great target of the quality of my work for building the rakugo speech database;

- All the former colleagues at HOYA Service Corp. and HOYA Corp. Various experiences with them supported my work and my life as a Ph.D. student;

- RevComm Inc., Tokyo, Japan, my affiliation as an employee since March 2020. Its flexible work arrangements allowed me to spare time for my Ph.D. study;

- All the colleagues at RevComm Inc. They always greatly encouraged me;

- Anyone who discussed rakugo speech synthesis with me. Their informative comments developed the quality of my work;

- Dr. Sojiro MAKINO, M.D., who is the Director of Makino Clinic, Tokyo, Japan, and was my doctor from August to November 2016. When I got a serious mental disorder, he first helped me as a doctor;

- Dr. Yojiro SAKAI, M.D., Ph.D., who is the Assistant Director of Akasaka Clinic, Tokyo, Japan, and my doctor since November 2016. He has been helped me as a doctor to deal with mental disorders;

- Ms. Maki IWASAKI (KATO), who is my wife. She has always supported and helped me mentally and financially even when I was in bed because of the serious mental disorder for about one year and five months. She also supported and helped me to recover from the disorder and live my life normally through her various efforts. She encouraged me so many times. She was next to me when I was generating the idea of my Ph.D. study (at the time, she was generating ideas of having a happy marriage life). She always helped me with practicing presentation, frequently discussed my work with me, and gave me a lot of important advice;

- Sasuke, who is our family's pet rabbit. He is always so cute and so relaxed that I could received a lot of energy from him and have a good rest like him.

I also would like to express my gratitude to:

- The Graduate University for Advanced Studies, SOKENDAI, my affiliation as a Ph.D. student. I am really lucky to find such a wonderful university;
- National Institute of Informatics, my affiliation as a Research Assistant and where I physically spent time during Ph.D. course. NII and its staffs kindly supported my work in various aspects.

Finally, I would like to express my appreciation to all the people who helped me to finish writing this thesis from just an idea jotted down on a paper.

修士論文テーマ

2016-6-16.

- 聞いていて安心するTTS.
- 漫才をするTTS.
- 動物の鳴き声をつくる.
- ものまねTTS.
- 落語をするTTS.

Figure 1: This thesis started from just an idea jotted down on a paper.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

The process in which a machine synthesizes speech is called speech synthesis. Especially in the case of converting an input text to speech, it is called text-to-speech (TTS). TTS is widely used in commercial products and services such as interactive voice response [7, 8, 9, 10, 11], car navigation systems [12, 13], voice assistants [14, 15, 16, 17], conversational systems for telephone [18, 19], reading books [20], screen readers [21, 22, 23, 24, 25, 26], broadcasting including narration for TV programs [27, 28], and announcements for public transport.

For decades, TTS could produce speech that sounds as intelligible but not as natural as human speech. Today machines have equipped the ability to synthesize speech that sounds as natural as human speech, albeit under limited conditions. Some of the TTS systems can produce speech with the same mean opinion scores (MOSs) as human speech [29, 30]. These systems are trained with well-articulated read speech. Researchers have made their endeavors to reach the naturalness of synthesized speech at the equivalent level of human speech for a long period, and the invention of end-to-end/sequence-to-sequence speech synthesis models following [31, 32, 33, 34] and neural vocoders such as [35] and succeeding ones finally enabled the realization of human-comparable TTS. Of course, attempts to model speech with various styles have also been actively investigated. Various methods have been proposed for representing speaking styles in end-to-end/sequence-to-sequence TTS [36, 37, 38, 39, 40, 41, 42, 43, 44, 45].

Then, are machines already equipped with enough ability to synthesize speech? The answer is no. One obvious example is that synthesized speech is deficient in stirring listeners' emotions, which is a role of speech other than just transferring information, such as content, speakers' emotion, personality, and intention, to listeners as a kind of media. In verbal entertainment, including *rakugo*, which is a traditional Japanese form of verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling, (human) performers entertain audiences through the medium of speech. Unfortunately, however, most of us could easily imagine and would agree that verbal entertainment performances by machines are quite unnatural or monotonic even if the content is appropriate. Although you might enjoy some speech-synthesis-based rakugo performances submitted to to online video platforms, mostly with many manual inventions [46, 47, 48], in our opinion, the quality of these performances is far inferior to that of performances by professional rakugo performers. We believe that such a gap between machines and humans should be filled to develop the human-machine relationship further. As an approach to this goal, we decided to start a challenge to build rakugo TTS that entertains audiences.

1.2 Thesis Overview

1.2.1 Motivation

As mentioned above, some of the current TTS systems can synthesize speech that sounds as natural as human speech in the case of read speech, but no TTS system can stir listeners' emotion enough, unlike humans, such as in the case of verbal entertainment including rakugo. We believe that such the gap between machines and humans should be filled. To evolve TTS in this aspect, we decided to develop rakugo

TTS as the first attempt to realize fully-fledged entertaining TTS.

1.2.2 Issues to Be Addressed

In this thesis, we focus on building rakugo speech synthesis that entertains audiences. As mentioned above, this is the first attempt to build full-fledged entertaining TTS. To realize such TTS, we have to solve some issues which we usually do not face when modeling speech ordinarily used for building speech synthesis:

Issue 1: There is no usable rakugo speech databases for speech synthesis.

Although thousands of commercial recording of rakugo speech are available, almost all of them contain noise and reverberation because of the recording environments (live recordings). Such speech is not suitable for building speech synthesis. We therefore have to prepare a speech database from scratch.

Issue 2: Rakugo speech is far more diverse and casually-pronounced than speech ordinarily used for building speech synthesis.

As described later, the main part of a rakugo story basically consists of conversations or dialogues between characters played by a performer, and there is little narrative speech. In other words, rakugo performers should tell whole stories without explicit explanations; therefore, rakugo speech in such conversational parts has truly varying speaking styles. Also, rakugo speech is more casually pronounced than speech ordinarily used for building speech synthesis because rakugo is performed improvisationally or from memory. Moreover, the Japanese language used in traditional rakugo stories, on which we focus in this thesis, is somewhat old-fashioned, and each character speaks a different dialect, sociolect, and idiolect according to his or her gender, age, social rank, and individuality. This diversity and casualness make it more difficult to properly train a model than the case of ordinary speech.

Issue 3: Characters should be easily distinguishable and contents should be easily understandable.

As mentioned in issue 2, the main part of a rakugo story basically consists of conversations or dialogues between characters played by a performer; therefore it is desired that listeners can easily distinguish characters and understand the contents when they listen to synthesized rakugo speech.

Issue 4: Synthesized speech should entertain listeners.

Since rakugo TTS we develop is an *entertaining* TTS, we have to measure not only the naturalness of speech it produces, which is usually measured in the evaluation of ordinary speech synthesis, but also whether it entertains listeners, and how well it does so. Because there is no established methodology to measure these aspects, we should develop such a methodology.

1.2.3 Contribution

On the issues above, this thesis makes the following contributions:

On issue 1:

We built the first rakugo speech database usable for TTS described in Chapter 3. We not only transcribed the speech but also appended context labels manually for each sentence.

On issue 2:

We successfully modeled rakugo speech using end-to-end/sequence-to-sequence TTS. As explained in Chapter 5, we attempted to model rakugo speech with the SSNT-based model, which has no soft attention network, to aim to deal with the diversity and casualness of rakugo speech. We also used GSTs or manually labeled context features to enrich speaking styles of rakugo speech. The models synthesized somewhat natural, character-distinguishable, and content-understandable speech. However, the MOSs for the speech were just around 3 through a listening test. For further improvement, we replaced the SSNT-based model above to Tacotron 2, which is a state-of-the-art TTS model, or an enhanced version of it with self-attention (SA-Tacotron) to better consider long-term dependencies explored in Chapter 6, f. We confirmed that the models synthesized speech of better quality than the SSNT-based models.

On issue 3:

As shown in Chapter 6, through a listening test, we found that state-of-the-art

TTS models could not yet reach the professional level, and there were statistically significant differences in terms of naturalness, distinguishability of characters, understandability of content, and even the degree of entertainment; nevertheless, the results of the listening test provided some interesting insights: 1) we should not focus only on naturalness of synthesized speech but also the distinguishability of characters and the understandability of the content to further entertain listeners; 2) the f_o expressivity of synthesized speech is poorer than that of human speech, and more entertaining speech should have richer f_o expression.

On issue 4:

As described in Chapter 7, we proposed a novel methodology for evaluating rakugo speech and conducted a listening test to investigate how the level of rakugo speech synthesis compares to professional rakugo performers at various levels. From the listening test results, we found that the level of speech synthesis did not reach that of human professionals; nevertheless, the results suggested that we should make the f_o expression of speech synthesis richer to better entertain audiences. This suggestion strengthens the finding of for Issue 3.

1.3 Outline of Thesis

In Chapter 2, we introduce rakugo and its performance. Since rakugo is not so famous outside Japan, we need to explain it in detail for your better understanding of this thesis.

In Chapter 3, we describe the details of our rakugo speech database. The speech was newly recorded, and the author transcribed it and labeled context features for each sentence. The database is designed to be suitable for building rakugo TTS. Statistical analysis for this database is also shown.

In Chapter 4, we give an overview of speech synthesis itself and end-to-end/sequence-to-sequence TTS, and describes audiobook TTS, which is an active research area and most similar to rakugo TTS, and the difference between audiobook speech/TTS and those of rakugo. We also refer to the relationship between speech synthesis and entertainment.

In Chapters 5 and 6, we introduce our experiment for building and evaluating rakugo

TTS. In Chapter 5, we describe an initial experiment to train sequence-to-sequence rakugo TTS models and evaluate them. In Chapter 6, we report a more sophisticated experiment.

In Chapter 7, we propose a novel methodology for evaluating rakugo speech and compare our rakugo synthesizer to human professionals.

In Chapter 8, we conclude with our findings, contributions, and future work.

2 Rakugo

2.1 Overview

2.1.1 History

Rakugo is a traditional Japanese form of one-person verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling. Although its history is unclear, it is said that professional rakugo performers, *hanashika*s, appeared in *Genroku* era (1688–1704) [49]. Tsuyu-no-Gorobe (–1703), Yonezawa Hikohachi, and Shikano Buzaemon (1649–1699) independently started their professional careers in Kyoto, Osaka, and *Edo* (Tokyo), respectively. In *Kamigata* (Kyoto and Osaka), performers performed outside, and their performances were called *tsujibanashi* (performance on the street). In Edo, performers performed inside, and their performances were called *zashikibanashi* (performance in a room).

In the late 1700s, theaters that mainly produce rakugo, *yose*s, appeared. From that time to early 1900s, rakugo became one of the most popular entertainment genres among ordinary people, and there were hundreds of yoses in Edo at the peak.

In *Meiji* era (1868–1912), rakugo started to be recorded in a book (*sokkibon*; shorthand book) or on a phonograph record. In 1925, radio broadcasting started in Japan, and rakugo also started to be broadcast via radio. This enabled people who could not go to yoses to enjoy rakugo performances. Even though the popularity of rakugo was declined because of the popularity of talkie and other new forms of entertainment, and many yoses were burned down by air raids during World War II, rakugo has been professionally performed until today.

2.1.2 Popularity

Although rakugo is a traditional form of entertainment with more than a 300-year history, it is popular even today in Japan. In Tokyo, there are four major yoses, and rakugo is performed in each one every day of the year, even on January 1 (Figure 2.1). There are many other minor yoses. Rakugo is also performed at small to large halls, restaurants, coffeehouses, bookstores, shrines, temples, etc. almost every day. Thousands of CDs, DVDs, and streaming audio/videos of rakugo performances by present or former professional rakugo performers are available. Some TV and radio programs are broadcast every week in Japan [50, 51, 52, 53, 54, 55, 56]. Amateur performances are also active. Some amateur rakugo performance societies at universities have produced professional performers.

2.1.3 Ranks of Professional Performers

Rakugo is generally divided into Edo (Tokyo) rakugo, on which we focus on in this thesis, and Kamigata rakugo, which has been developed in Osaka and Kyoto. A professional rakugo performer is called a hanashika, as mentioned earlier. In Edo rakugo, a hanashika is ranked at one of three levels, i.e., *zenza* (minor performer, assistant of stages, and housekeeper at their master or mistress's house), *futatsume* (second-rank performer), and *shin-uchi* (first-rank performer). Only shin-uchis can take disciples. Usually, it takes about 3 to 5 years to be promoted from zenza to futatsume, and about 10 years to be promoted from futatsume to shin-uchi. About 600 performers are active as professionals in Edo rakugo as of 2020.

Figure 2.1: Four major yoses (theaters mainly producing rakugo) in Tokyo. Upper left: Suzumoto Engeijo [1], upper right: Suehirotei [2], lower left: Asakusa Engei Hall [3], lower right: Ikebukuro Engeijo [4].

2.1.4 Kinds of Stories

Rakugo is derived from telling a very short comic story, called *kobanashi*. A kobanashi has an *ochi*, the punch line, at the end of the story. It evolved into a longer and more complex story, called *otoshibanashi* (a story with ochi) or *kokkeibanashi* (a comic story), which concentrates on making audiences laugh. In later times, a story without ochi started to be performed. It is called *ninjobanashi* (a story with humanity), which concentrates on impressing audiences. In this thesis, we focus on otoshibanashi (kokkeibanashi).

2.1.5 Structure of a Story

A rakugo story is composed of five parts: *maeoki* (greeting), *makura* (introduction), the main part, ochi (punch line), and *musubi* (conclusion) [57]. Maeoki is optional, so it may not appear during a performance. As mentioned above, some stories such

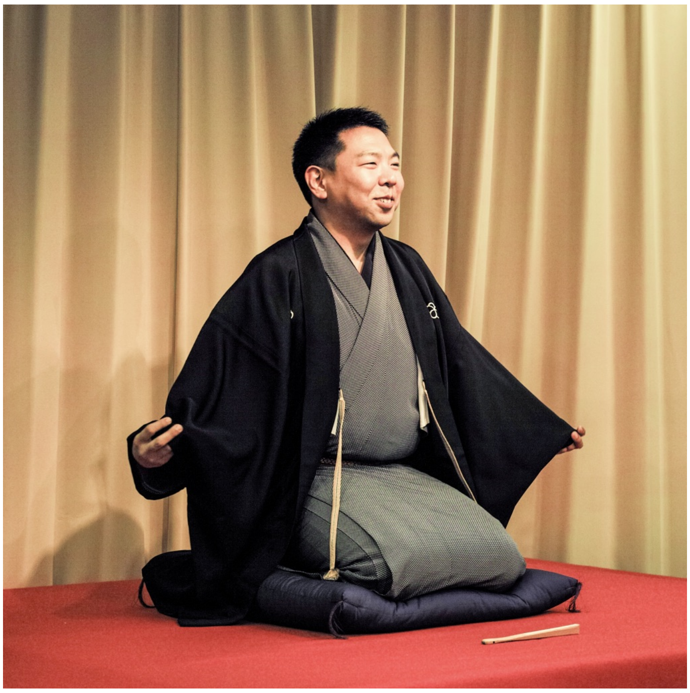

Figure 2.2: Shumputei Shotaro [5], who is a professional rakugo performer, performing rakugo on a stage [6].

as ninjobanashi do not have ochi, and have musubi in place of ochi. Musubi is also used when performers terminate stories because of time limitations. Makura is often improvised, but during this, performers basically do not have conversations with the audience, unlike stand-up comedy. Ochi, the punch line, is the most important part of rakugo (the word "rakugo"（落語）is derived from "a story with ochi"（落ち）).

2.2 Performance

During a performance, a rakugo performer sits down on a *zabuton* (cushion) and performs improvisationally or from memory alone on a stage (Figure 2.2). He or she plays multiple characters, and their conversations and dialogues make the story progress. In the main part of a story, to be mentioned later, almost all of the parts consist of conversations and dialogues between the characters played by the performer. In Edo rakugo, performers use only a *sensu* (folding fan) and a *tenugui* (hand towel) as props.

Rakugo stories are generally divided into standards, which were established by about the 1920s, and modern stories, which were created after the 1930s. In this thesis, we focus on standards. It should be noted that the Japanese language used in standards is slightly old-fashioned, and each character speaks a different dialect, sociolect, and idiolect of Japanese according to his or her gender, age, social rank, and individuality.

The length of rakugo stories varies from story to story. Even if performers play the same story, the length can be varied from stage to stage because of time limitations or other situations. In a yose, one hanashika usually performs for about 15 minutes (only the last performer performs for about 30 minutes). In other stages or recordings, they may perform longer.

Rakugo stories are taught through oral instruction from a master or mistress to a disciple except when the story is new. Performers may edit stories to increase the quality or match their own characteristics. They sometimes insert jokes not only in the makura but also in the main parts of the stories according to the situations during their performances.

The following is an example of a very short rakugo paragraph (otoshibanashi).

> Tome: Whoa! Oh no! Oh no! Oh no! Oh no!
> Friend: Wait Tome. What are you doing?
> Tome: Oh, I'm chasing after a thief.
> Friend: Seriously? Aren't you the fastest man in this town? He is unlucky.
> Friend: Which direction did he escape?
> Tome: He's catching up with me.

2.3 Academic Research on Rakugo

A lot of academic research on rakugo has been carried out in many fields such as literature, linguistics, phonetics, history, education, and psychology. Rakugo is also used as a subject to some engineering study [58, 59]. However, rakugo has never been a subject to speech processing study even though rakugo indeed consists of speech. This research is the first authentic rakugo speech synthesis study.

3
Database

3.1 Motivation

We built a large rakugo speech database for our study because there were no rakugo speech databases suited to speech synthesis. Most commercial rakugo recordings are live recordings that include noise and reverberation; therefore, we recorded the rakugo speech ourselves.

3.2 Recording

3.2.1 About the database

We built two sub-databases. One is used for training speech synthesizers, described in Chapters 5 and 6, and the other is used exclusively for evaluation via a listening test, described in Chapter 7. We call the first sub-database “Database I” and the second one “Database II.”

Figure 3.1: Yanagiya Sanza performing rakugo alone in recording booth.

3.2.2 Database I

The recordings were conducted from July to September 2017. The performer was Yanagiya Sanza [60], a professional rakugo performer with over 20 years of experience and who was promoted to shin-uchi in 2006. Only he was in the recording booth, and he did not face or receive any reaction from an audience unlike a real performance on a stage (Figure 3.1). This environment is unusual as a performance, but we think it is reasonable in a sense because performers including him definitely often practice alone and he could perform as he liked. He performed 25 Edo rakugo standards, lasting from 6 to 47 minutes length (total 13.2 hours including pauses between sentences) (Table 3.1). We did not re-record any of the performance because of mispronunciation or restatements except in cases where the performer asked us to do so, because he said that the flow of the performance is very important for performing rakugo.

Table 3.1: Details of recorded stories in Database I.

Name	Dur. (mm:ss)	Rec. date	Annotated	Characters in makura	Characters in main part
道灌 (*Dokan*)	21:16	Jul 26, 2017	✓	-	*Hachigoro*, retired man, young man, Hachigoro (acting woman)
五貫裁き (*Gokansabaki*)	43:17	Jul 26, 2017	✓	*Mito Komon*, *Kaku-san*, *Toyama-no-Kin-san*	*Hachigoro*, *Tarobe*, *Ooka Echizen-no-kami*, all the people there, *Tokuriki-ya Man-emon*, *Sakube*, *Sadajiro*, five town officials, thief-taker, police officer
真田小僧 (*Sanadakozo*)	30:59	Jul 26, 2017	✓	child #1, child#2, mother, child#3	father, *Kimbo*, mother
藁人形 (*Waraningyo*)	38:29	Jul 26, 2017		cat flea remove contractor, customer #1, ear cleaner, customer #2	*Sainen*, *Okuma*, *Kisuke*, *Jinkichi*, caretaker
粗忽の使者 (*Sokotsu No Shisha*)	32:00	Jul 26, 2017		office worker, shop owner, *Sadakichi*	*Sugidaira Masame-no-sho*, *Jibuda Jibuemon*, vassal #1, *Akai Gomon-no-kami*, groom #1, groom #2, carpenter, *Tomekko*, vassal #2, *Tanaka Sandayu*
金明竹 (*Kimmeichiku*)	30:28	Jul 26, 2017	✓	younger brother, elder brother, father, mother	shop owner, *Matsuko*, wife, man, *Omiya*, western young man
転失気 (*Tenshiki*)	21:59	Aug 15, 2017		guide, madam	doctor, priest, *Chinnen*, flower seller, flower sellor's wife, doctor's assistant
ろくろっ首 (*Rokurokkubi*)	33:34	Aug 15, 2017	✓	-	ancle, *Matsuko*, nanny
粗忽の釘 (*Sokotsu No Kugi*)	37:28	Aug 15, 2017		son, father	wife, husband, neighbor across the street, neighbor
田能久 (*Tanokyu*)	30:58	Aug 15, 2017		-	*Kyube*, member of Kyube's theater company, *Guts Ishimatsu*, lumberjack, anaconda, Kyube (acting young woman), villager
青菜 (*Aona*)	37:26	Aug 15, 2017	✓	boy, girl, pupil, beautiful woman	retired man, gardener, retired man's wife, young man, gardener's wife, carpenter
蒟蒻問答 (*Konnyakumondo*)	36:07	Aug 15, 2017	✓	customer, artisan #1, artisan #2	*Rokube*, *Hachigoro*, *Gonsuke*, *Shamitakuzen*
権助提灯 (*Gonsukechochin*)	17:04	Aug 15, 2017	✓	Yama, office worker, husband #1, young man #1, young man #2, maid, madam, husband #2	Wife, husband, *Gonsuke*, mistress
やかん (*Yakan*)	20:10	Sep 20, 2017		-	*Hachigoro*, teacher
道具屋 (*Doguya*)	28:33	Sep 20, 2017	✓	younger brother, elder brother, father, mother	*Mokube*, *Yotaro*, police officer, *Tomozo*, carpenter, rickshaw driver, elderly man, vigorous man
寝床 (*Nedoko*)	40:30	Sep 20, 2017	✓	rakugo pupil, rakugo master, *gidayu* pupil, gidayu master	shop owner, *Shigezo*, servant, head clerk, guest #1, guest #2, *Sadakichi*
元犬 (*Motoinu*)	31:47	Sep 20, 2017	✓	child, young man #1, *Kin-chan*, mother, father, young man #2, young man #3, elderly woman, *Hanasaka-jisan*, dog *Pochi*, elderly man, young woman, young man #4, dog *Kuro*	visitor #1, visitor #2, visitor #3, white dog, *Kazusa-ya Kichibe*, retired man, *Omoto*
大工調べ (*Daikushirabe*)	43:47	Sep 20, 2017	✓	carpenter	*Masagoro*, *Yotaro*, *Genroku*, gatekeeper, magistrate, townsperson
転宅 (*Tentaku*)	23:52	Sep 20, 2017		*Tomi*, young man	*Okiku*, shop owner, *Saigobe*, tobacconist
うどん屋 (*Udon-ya*)	33:38	Sep 22, 2017		ragman #1, young man #1, young man#2, novice ragman, young man #3, expert ragman, ragmen #4, young man #4	young man #1, young man #2, young man #3, *soba* vendor, *udon* vendor, drunk, mother, customer
お菊の皿 (*Okiku No Sara*)	34:38	Sep 22, 2017	✓	young man, child(ren), audience #1, teacher, zenza, audience #2, ghost character, child audience, ghost, ghost speaking Osaka dialect	young man, retired man, villager, *Aoyama Tessan*, *Kiku*, brave young man, timid young man, new audience #1, regular audience #1, new audience #2, regular audience #2
締め込み (*Shimekomi*)	30:23	Sep 22, 2017		*Nio*, thief	thief, *Hachigoro*, *Ofuku*
茶の湯 (*Chanoyu*)	47:10	Sep 22, 2017	✓	-	retired man, *Sadakichi*, *tofu* maker/seller, tofu maker/seller's wife, head-firefighter, teacher, neighbor middle-aged man #1, neighbor middle-aged man #2, friend, farmer
お血脈 (*Okechimyaku*)	39:24	Sep 22, 2017	✓	(no makura)	audience #1, young man, retired man, magician, hanashika, audience #2, neighbor, middle-aged man, Buddha, Amitābha, master, zenza, *Moriya-no-otodo*, Prime minister *Mori Yoshiro*, Prime minister's staff, President Bill Clinton, Amitābha (platinum body), *Honda Yoshimitsu*, audience #3, Yama, *Miru-me-kagu-hana*, hell staff, *Ishikawa Goemon*, Guts Ishimatsu
味噌豆 (*Misomame*)	6:45	Sep 22, 2017	✓	young man #1, young man #2, crab, elderly husband, elderly wife	shop owner, *Sadakichi*

3.2.3 Database II

The recordings were conducted in January 2020. In these recordings, we recorded the performances of a common story told by professional performers of the three ranks. The performers were Yanagiya Kogoto (zenza), Ryutei Ichido (futatsume), and Yanagiya Sanza, who performed for the Database I. The recording conditions were the same as those of the Database I. The story performed by them is called "Misomame." The total durations of the recordings by the zenza, futatsume, and shin-uchi were 2.5, 2.7, and 4.2 minutes, respectively.

It should be noted that specific wording depends on performers because rakugo stories do not have any scripts. The shin-uchi performer attended the recording session of the zenza performer, and he supervised and instructed the zenza performer when necessary. This is because the zenza performer did not have the skill to perform the story as his routine unlike the shin-uchi and futatsume performers.

3.3 Annotation and Processing

3.3.1 Transcription

The author carefully transcribed the pronunciation of the recorded speech in the Database I. We did not define any special symbols for mispronunciation, fillers, or laughs. We used a comma only at a pause in a sentence, a period at the end of a sentence, and a question mark at the end of a sentence that ends with rising intonation. The ratio of question sentences, those having a question mark at the end, to the other sentences is about 3:7. We separated sentences according to the following criteria.

- A place we can separate sentences grammatically followed by a pause.
- A place where a turn-taking occurs.
- A place right after a rising intonation.

All the symbols used in the transcription are listed in Table 3.2. We did not use any accent symbols, although Japanese is a pitch-accent language, because the results of automatic morphological analysis and accent estimation are not usable because of

Table 3.2: Symbols used in transcription of Database I.

Phonemes	Vowels	a, e, i, o, u
	Consonants	b, by, ch, d, dy, f, fy, g, gw, gy, h, hy, j, k, kw, ky, m, my, n, N, ng, ny, p, py, r, ry, s, sh, t, ts, ty, v, w, y, z
	Other	cl (geminate consonant)
Pauses		pau (comma), sil (start of a sentence and period), qsil (question mark)

the difference between the slightly old-fashioned Japanese dialects, which is spoken by characters in the stories, and the modern standard Japanese. Of course, manual labeling of accents is impractical because it is time consuming.

3.3.2 Context Labels

We also appended context labels to each sentence in the Database I (Table 3.3). All the labels, excluding **part**, were defined by the author because no well-known categories of them exists in rakugo.

We believe the **role** of the character is important because almost all speech in rakugo, especially in the main part, is composed of conversations or dialogues between multiple characters. The **individuality** of the character is a special category for fool characters, usually called *Yotaro*, who often appear in rakugo stories. We believe the **condition** of characters is also important because characters speak in various styles according to their emotions, intention, the situations, etc. All the styles were defined by the author via carefully listening to speech and reading context. The **relationship** of the talking companions was defined because in conversations or dialogues (between two characters) in rakugo, one must be considered the superior and the other as an inferior. The **n_companion** (number of the talking companions) was defined because characters may talk to themselves or speak to one person or multiple persons. The **distance** to the talking companions was defined because characters may speak to someone near or far from them. In the context of a particular **part** of the story, we considered maeoki (greetings) and musubi (conclusion) as makura (introduction) and ochi (punch line), respectively.

Table 3.3: Context labels (*hanashika* refers to speech not by any characters).

Group	Name	Details
ATTRibute of character	**role** of character	**gender**: *hanashika*, male, female; **age**: *hanashika*, child, young, middle-aged, old; **social rank**: *hanashika*, *samurai*, artisan, merchant, other townsperson, countryperson, with other dialect, modern, other
	individuality of character	*hanashika*, not fool, fool
CONDition of character	**condition** of character	neutral, admiring, admonishing, affected, angry, begging, buttering up, cheerful, complaining, confident, confused, convinced, crying, depressed, drinking, drunk, eating, encouraging, excited, feeling sick, feeling sorry, feeling suspicious, finding it easier than expected, freezing, frustrated, ghostly, happy, hesitating, interested, justifying, *kakegoe* (shouting/calling), loud voice, laughing, leaning on someone, lecturing, looking down, panicked, pet-directed speech, playing dumb, putting up with, rebellious, refusing, sad, scared, seducing, shocked, shouting, sketchy, small voice, sleepy, soothing, straining, surprised, swaggering, teasing, telling off, tired, trying to remember, underestimating, unpleasant
SITuation of character	**relationship** to talking companions	*hanashika*, narrative, soliloquy, superior, inferior
	n_companion: number of talking companions	*hanashika*, narrative, soliloquy, one, two or more
	distance to talking companions	*hanashika*, narrative, near, middle, far
STRucture of story	**part** of story	makura, main part, ochi

3.4 Statistical Analysis on Acoustic Features

3.4.1 Overview

In order to investigate characteristics of rakugo speech, we performed some basic statistical analyses on acoustic features, i.e. mean logarithmic fundamental frequency ($\ln f_o$) and duration per mora, for each of the context labels using the Database I.

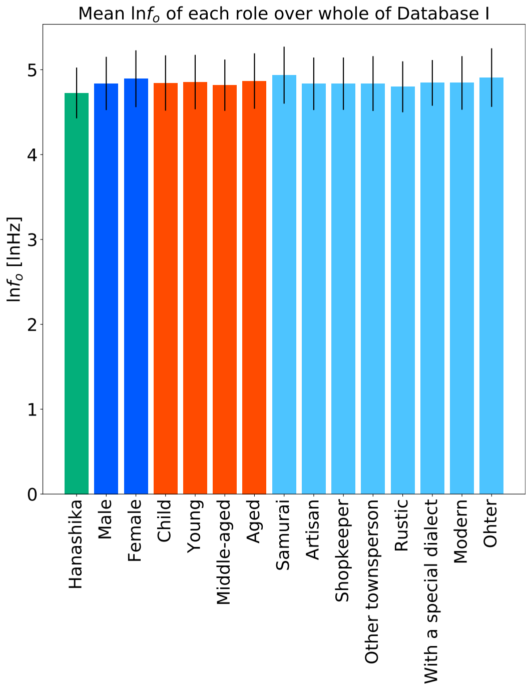

Figure 3.2: Mean lnf_o of each **role** over whole of Database I.

3.4.2 Role

Mean lnf_o and duration per mora of each **role** over the whole of Database I is shown in Figures 3.2 and 3.3. Green, blue, red, and light blue bars refer to hanashika, gender, age, and social rank, respectively. We can find that the performer (Yanagiya Sanza) did not obviously distinguish characters from each other except hanashika using f_o. On the other hand, he differentiates speaking rate (duration per mora) to distinguish characters. It should be noted that the manner to distinguish characters depends on performers (e.g. Chapter 7).

3.4.3 Individuality

Mean lnf_o and duration per mora of each **individuality** over the whole of Database I is shown in Figures 3.4 and 3.5. Similar to the case of **role**, the performer represents a character's individuality more by changing speaking rate than by changing f_o.

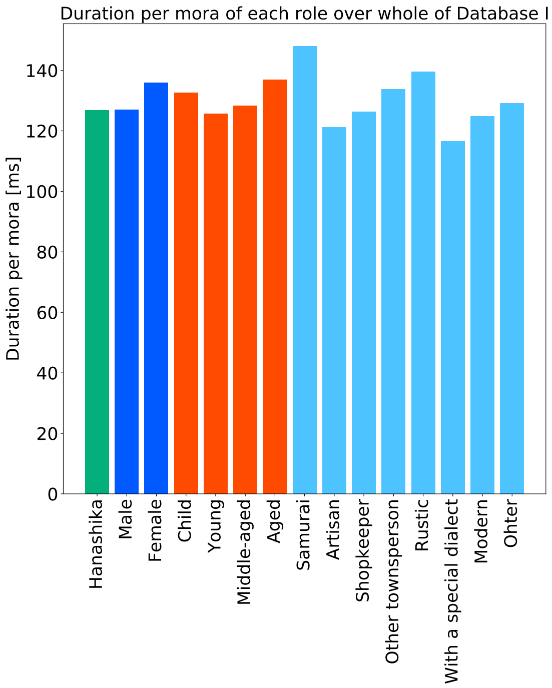

Figure 3.3: Duration per mora of each **role** over whole of Database I.

3.4.4 Condition

Mean lnf_o and duration per mora of each **condition** over the whole of Database I is shown in Figures 3.6 and 3.7. We can find that the performer uses both f_o and speaking rate to act various speaking styles. In particular, he drastically changes speaking rate for some **condition**s.

3.4.5 Relationship

Mean lnf_o and duration per mora of each **relationship** over the whole of Database I is shown in Figures 3.8 and 3.9. Similar to the case of **role**, the performer represents relationship more by changing speaking rate than by changing f_o.

3.4.6 N_companion

Mean lnf_o and duration per mora of each **n_companion** over the whole of Database I is shown in Figures 3.10 and 3.11. Similar to the case of **role**, the performer represents

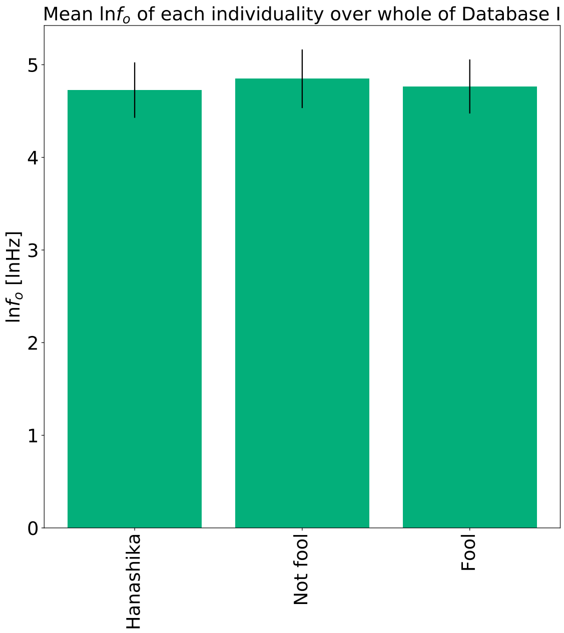

Figure 3.4: Mean lnf_o of each **individuality** over whole of Database I.

the number of talking companions more by changing speaking rate than by changing f_o.

3.4.7 Part

Mean lnf_o and duration per mora of each **part** over the whole of Database I is shown in Figures 3.12 and 3.13. Similar to the case of **role**, the performer differentiates speaking rate more than f_o for each **part**.

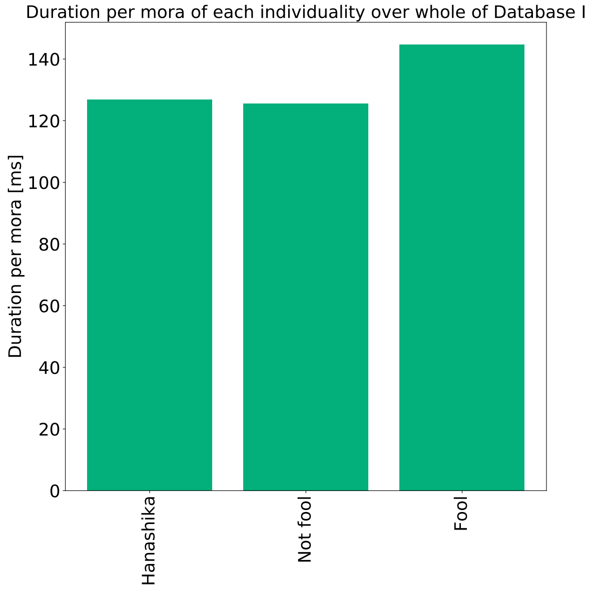

Figure 3.5: Duration per mora of each **individuality** over whole of Database I.

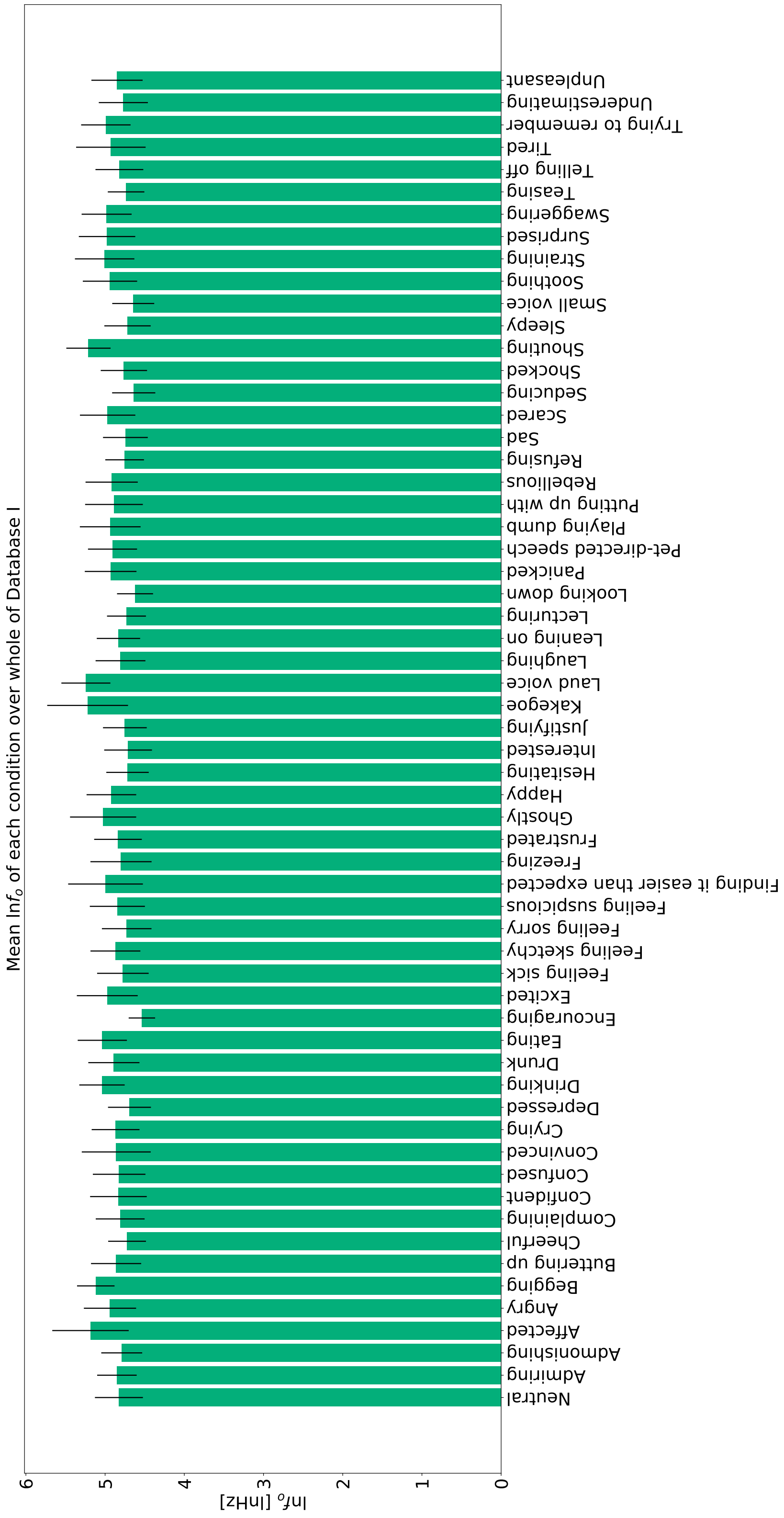

Figure 3.6: Mean lnf_o of each **condition** over whole of Database I.

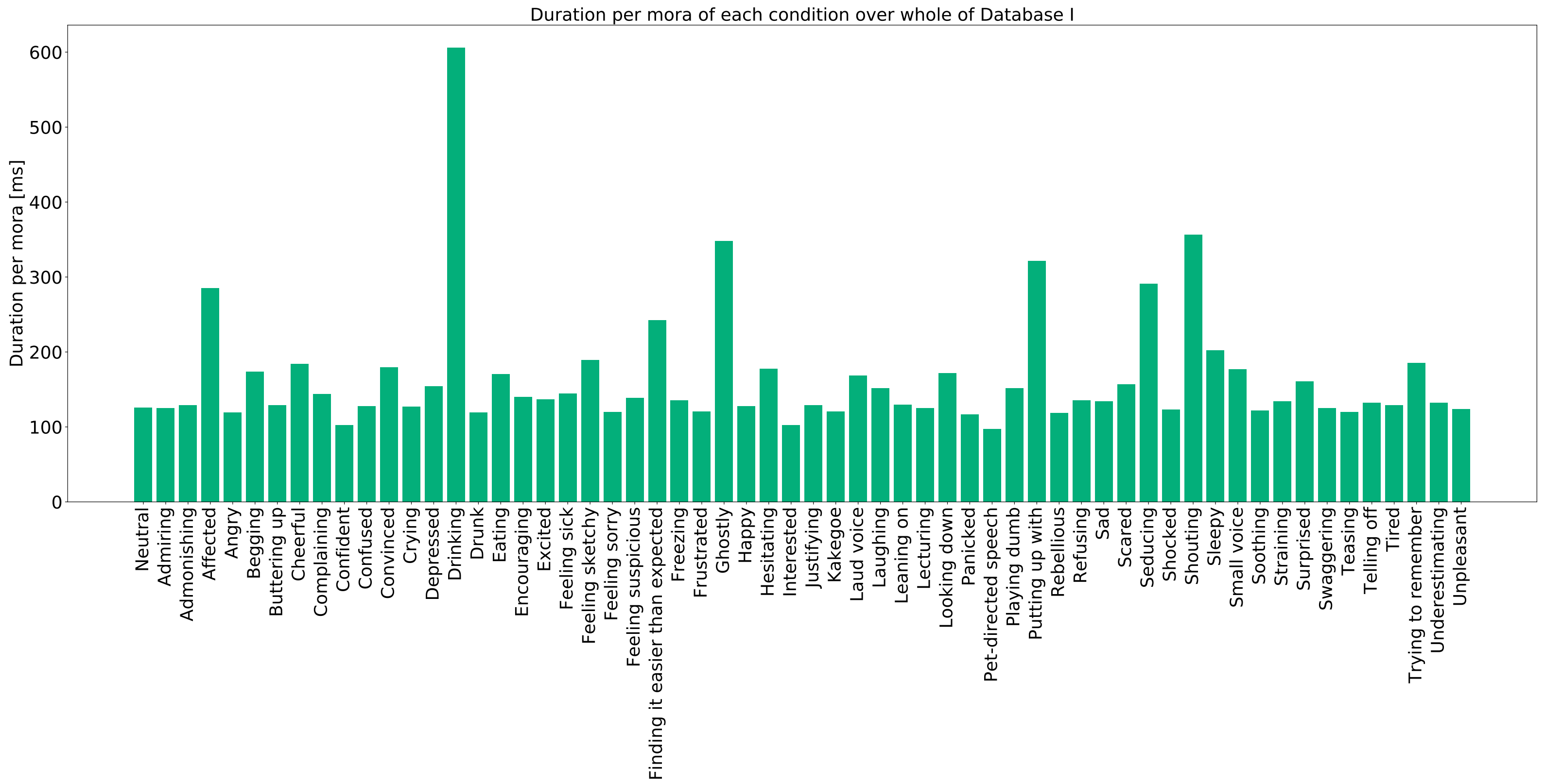

Figure 3.7: Duration per mora of each **condition** over whole of Database I.

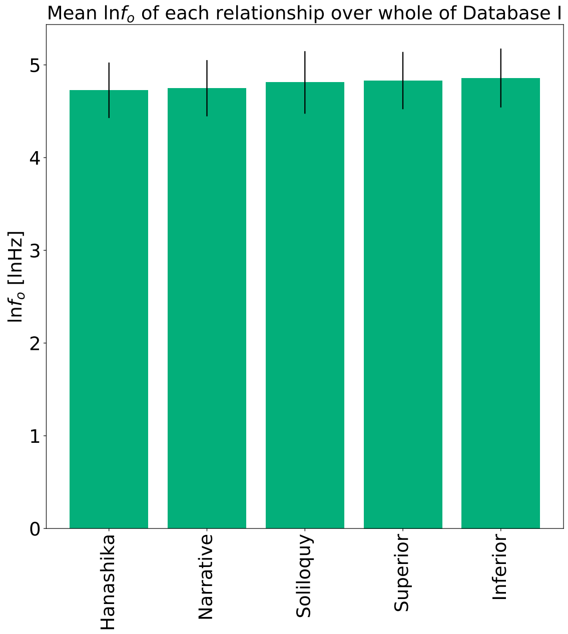

Figure 3.8: Mean lnf_o of each **relationship** over whole of Database I.

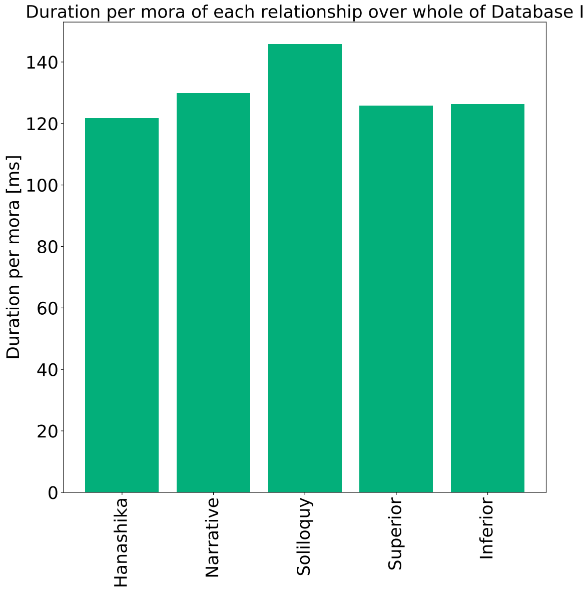

Figure 3.9: Duration per mora of each **relationship** over whole of Database I.

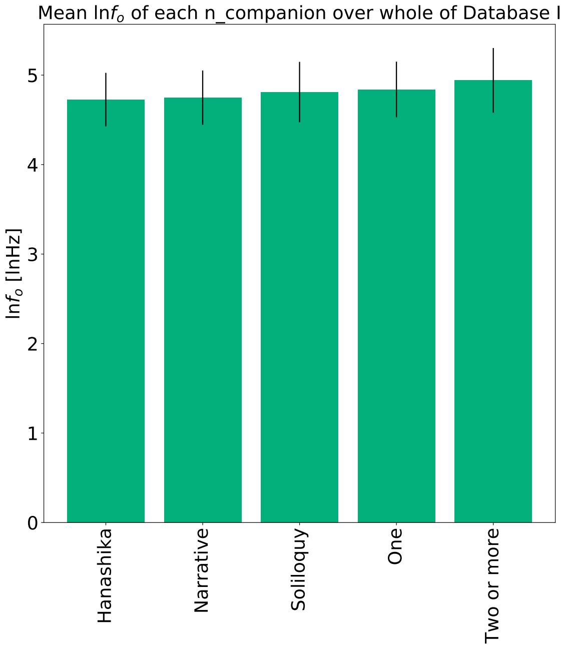

Figure 3.10: Mean lnf_o of each **n_companion** over whole of Database I.

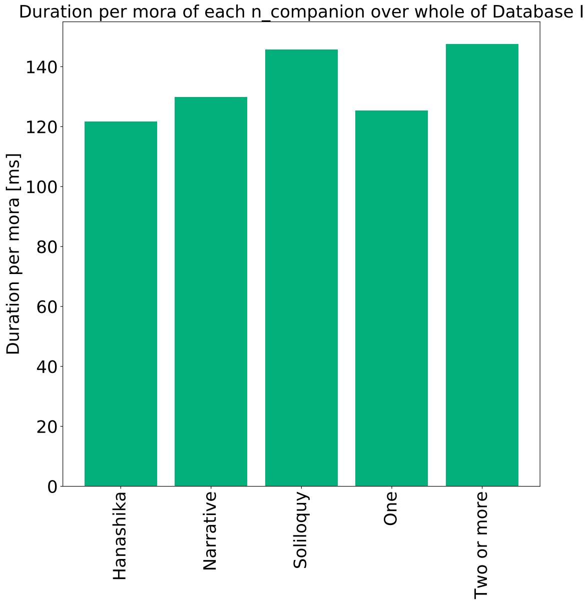

Figure 3.11: Duration per mora of each **n_companion** over whole of Database I.

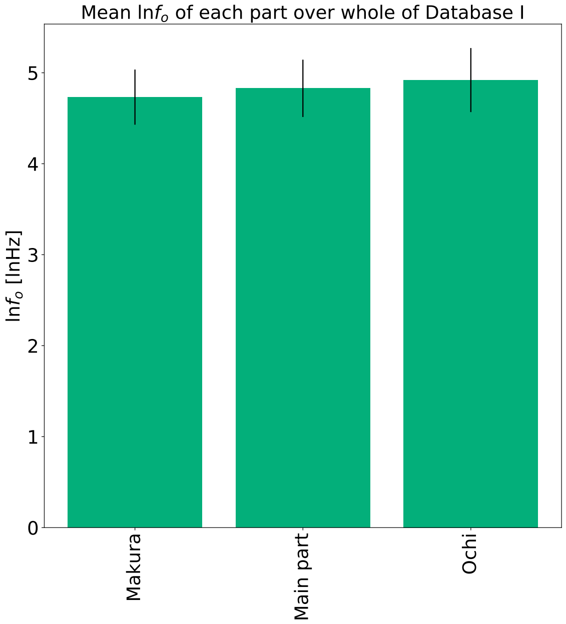

Figure 3.12: Mean lnf_o of each **part** over whole of Database I.

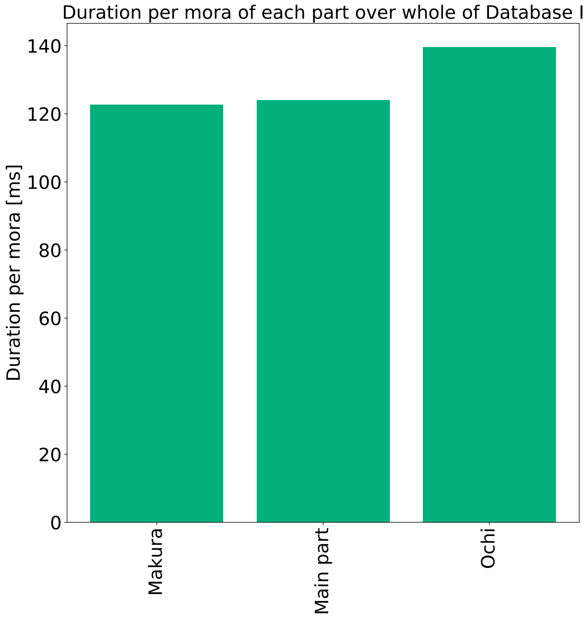

Figure 3.13: Duration per mora of each **part** over whole of Database I.

4

Speech Synthesis and Its Relationship to Entertainment

4.1 Introduction to Speech Synthesis

4.1.1 Overview

Speech synthesis is the process in which a machine produces speech. Speech synthesis converts source information to speech, and in current speech synthesis, the source information is typically a text (text-to-speech; TTS) (Figure 4.1). Speech synthesis has a history of hundreds of years [61].

Today, we face a lot of synthesized speech in our everyday lives. Speech synthesis is broadly used in commercial products and services such as interactive voice response [7, 8, 9, 10, 11], car navigation systems [12, 13], voice assistants [14, 15, 16, 17], conversational systems for telephone [18, 19], reading books [20], screen readers [21, 22, 23, 24, 25, 26], broadcasting including narration for TV programs [27, 28], and announcements for

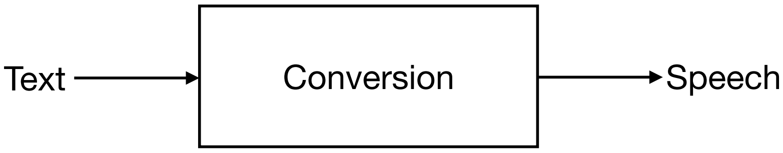

Figure 4.1: Example structure of TTS.

public transport. In such products and services, speech synthesis remarkably reduces human efforts and/or makes it possible to produce products or services that would be unrealistic with using only human's speech.

4.1.2 Concatenative Speech Synthesis vs. Statistical Parametric Speech Synthesis

Most practical TTS is computer-based, and there are two major methods: Concatenative speech synthesis (CSS) and statistical parametric speech synthesis (SPSS). CSS concatenates recorded waveforms or waveform units into an output waveform matching the input. Especially in the case of concatenating adaptively-split waveform units into a waveform matching the input of any content, it is called unit selection speech synthesis (USS). SPSS calculates an output waveform based on statistical parametric model(s) matching to the input. Hidden-Markov-model-based (HMM-based) models [62] and deep-learning-based ones [63] are famous and frequently used.

As mentioned above, CSS concatenates recorded waveforms or waveform units into an output waveform matching the input. Figure 4.2 shows the example mechanism of CSS (USS). In this mechanism, waveform units are concatenated considering target costs (how well the unit matches the input) and concatenation costs (how smooth the neighbor units are concatenated). CSS has been widely used commercially because it needs relatively light computational costs during synthesis, and it can produce speech as natural as human albeit under limited conditions, especially in the case that an output waveform that perfectly matches the entire input has been recorded and is stored in the database (in short, just playing the recorded speech). These features are reasonable and attractive in commercial use. However, the quality of synthesized speech can be easily decreased when no suitable waveforms exists in the database. Even worse, CSS is not flexible. For example, CSS is not good at synthesizing

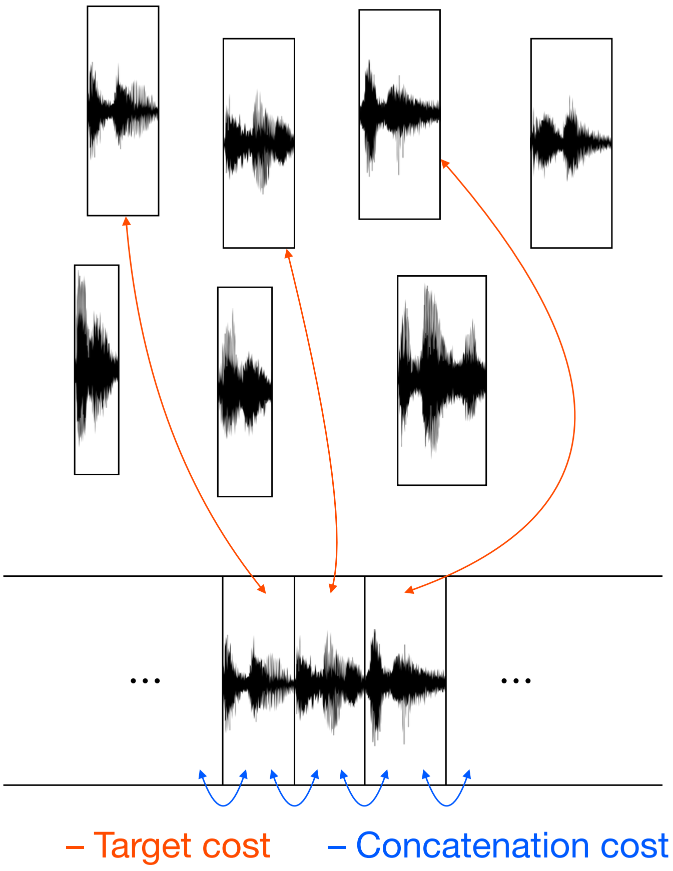

Figure 4.2: Example mechanism of CSS (USS).

a waveform in various speaking styles because it just concatenates waveform units to synthesize an output waveform. If we want to change speaking styles of CSS, we basically have to record enough speech and build an independent model from scratch for each speaking style. The same is true for changing speakers.

On the other hand, SPSS (Figure 4.3) is far more flexible. Since SPSS generates an output waveform based on statistical models, we do not have to build an independent model from scratch for each speaking style or speaker. We can adapt a model from an existing one, or just build a mixed model using all the speaking styles and speakers with their identifiers. Moreover, although HMM-based models could not produce speech that has human-comparative naturalness and they have suffered from their "synthetic" voice, which is derived from oversmoothing of parameters, some deep-learning-based models finally succeeded in producing speech that sounds as natural as human

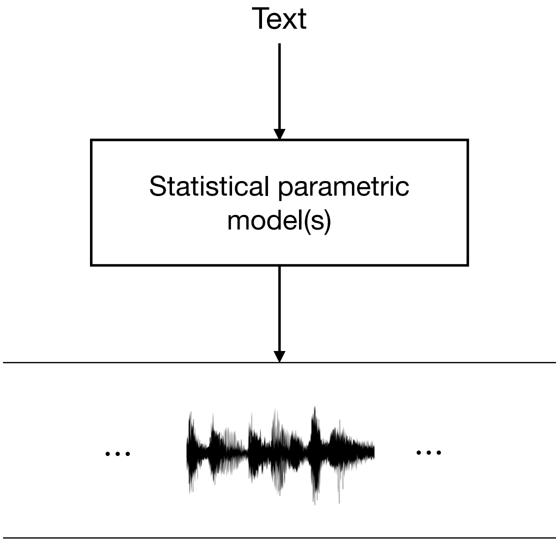

Figure 4.3: Example structure of SPSS.

speech [29, 30], albeit in the case of well-articulated read speech. Today SPSS is spreading its commercial use [64, 65] even though it needs relatively (HMM) to very high (deep learning) computational costs during synthesis.

4.1.3 Pipeline/Frame-by-Frame Model vs. End-to-End/Sequence-to-Sequence Model

Conventionally, it has been difficult to directly convert input text to output waveform; therefore the converting process has been usually split into some sub-processes to be solved independently. A TTS model based on such a framework is called a pipeline model (Figure 4.4). Because acoustic features in most pipeline SPSS are estimated frame by frame based on estimated phonetic duration, the term frame-by-frame model is therefore also frequently used. The typical process of pipeline/frame-by-frame models is as follows:

Sub-process 1: Convert input text to linguistic information by a text analyzer.

Sub-process 2: Convert the linguistic information to acoustic information by acoustic models.

Sub-process 3: Convert the acoustic information to output waveform by a (conventional signal-processing-based) vocoder.

In the sub-process 1, an independent text analyzer converts an input text to linguistic information, such as phonemes, accentual information, positional information, accentual phrase boundary, intonation phrase boundary, and parts of speech. In the sub-process 2, the duration of each phoneme is estimated, and spectral parameters and fundamental frequencies (f_os) are estimated from the linguistic information within the duration. Finally in the sub-process 3, the output waveform is generated through a (conventional signal-processing-based) vocoder such as [66, 67, 68, 69] from further estimated spectral parameters and f_os frame by frame.

Pipeline/frame-by-frame models have been widely used because of practical reasons and their ease of maintenance. For example, when we get a linguistic error during synthesis, we can just fix the text analyzer and do not need to fix the other modules. Such ease of maintenance is attractive especially in commercial use. However, because the sub-processes above are optimized independently, the output waveform is generated not considering the global optimization, and necessarily the quality tends to be not ideal.

To deal with this problem, many attempts have been actively conducted to combine the sub-processes using deep neural networks and get close to the global optimization. A TTS model based on such an approach is called an end-to-end model (Figure 4.5). Typical end-to-end models estimate an acoustic feature sequence based not on frame by frame but on an input text sequence, the term sequence-to-sequence model is therefore also frequently used. End-to-end/sequence-to-sequence models greatly improved the quality of synthesized speech, and they contributed to and enabled the realization of TTS that can produce speech that sounds as natural as human speech in the case of well-articulated read speech [29, 30].

4.1.4 Typical Structure of End-to-End/Sequence-to-Sequence TTS

While many end-to-end/sequence-to-sequence TTS models have been proposed, Tacotron 2 [29] and Transformer TTS [30] are representative ones. Both of them are classified as autoregressive end-to-end/sequence-to-sequence TTS. A typical structure of autoregressive end-to-end/sequence-to-sequence TTS is shown in Figure 4.5. An

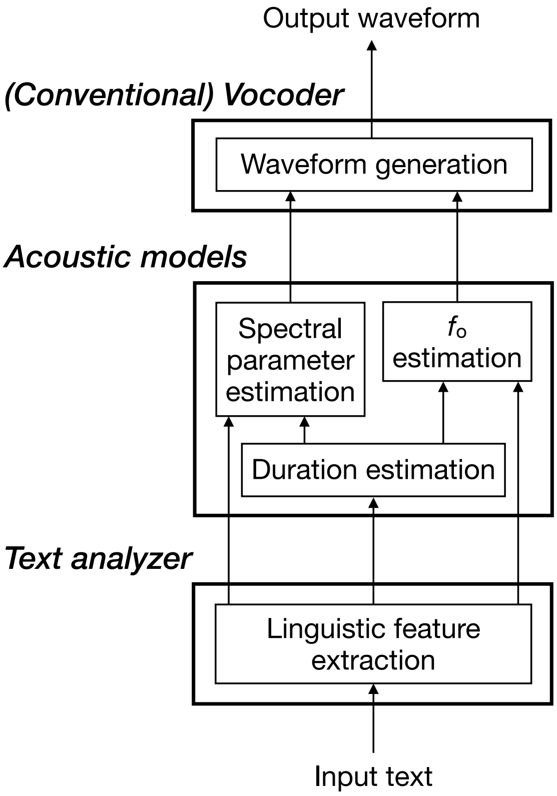

Figure 4.4: Example structure of pipeline/frame-by-frame TTS.

autoregressive end-to-end/sequence-to-sequence TTS has an encoder, decoder, and attention modules. These modules convert an input text sequence to a spectrogram. Then a neural vocoder is used to convert the spectrogram to a waveform.

Encoder

An encoder converts an input text sequence to a hidden encoded feature sequence. In the case of Tacotron 2, the encoder is composed of a stack of convolutional neural networks (CNNs) and a bi-directional long short-term memory (LSTM) [70]. In the case of Transformer TTS, the encoder is composed of an encoder pre-net, which is composed of CNNs, and a stack of encoder blocks, each of which is composed of a multi-head attention and a feed-forward network (FFN).

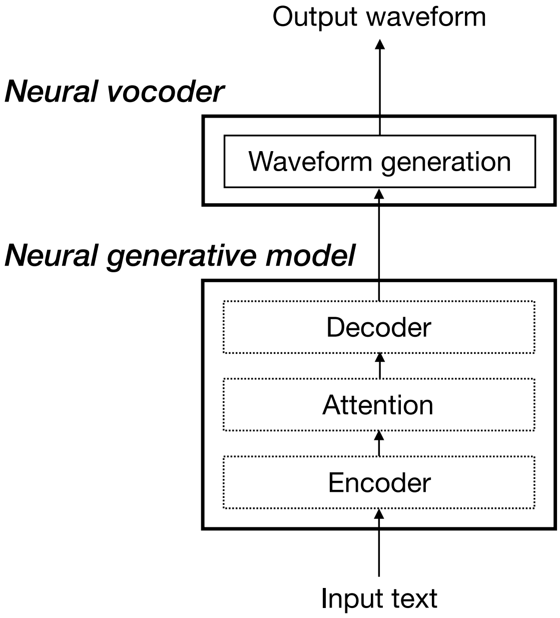

Figure 4.5: Example structure of end-to-end/sequence-to-sequence TTS.

Decoder

A decoder converts the hidden encoded feature sequence to a spectrogram. In the case of Tacotron 2, the decoder is composed of an FFN and LSTMs. In the case of Transformer TTS, the encoder is composed of an decoder pre-net, which is composed of FFNs, and a stack of decoder blocks, each of which is composed of a masked multi-head attention, a multi-head attention, and an FFN. Both in the cases of Tacotron 2 and Transformer TTS, the decoders predict a spectrum of a time step based on the spectrum of the previous time step. That is the reason they are called autoregressive end-to-end/sequence-to-sequence models.

Attention

When a decoder converts the hidden encoded feature sequence to a spectrogram, the decoder predicting a spectrum of a time step should pay attention to the related locations of the hidden encoded feature sequence. An attention module serves such a function. Namely, the attention module maps the encoder-decoder time steps.

Attention module is classified into two types: Soft attention and hard attention. Soft attention maps a time step of the decoder and time step*s* of the encoder where sum of the mapping weights equals to 1. On the other hand, hard attention maps a time step of the decoder and *a* time step of the encoder.

Neural Vocoder

A neural vocoder is a neural network that converts a spectrogram to a waveform. After the appearance of WaveNet [35] vocoder [71, 72] as the first neural vocoder, many neural vocoders have been proposed [73, 74, 75, 41, 76]. We use a WaveNet vocoder in this thesis because it is a state-of-the-art neural vocoder with regard to the quality of the output waveform while its inference speed is very slow.

Neural vocoders can produce waveforms with prominently better quality than conventional signal-processing-based vocoders can in terms of speech synthesis. Partially owing to their outstanding performance, as mentioned in 4.1.3, the combination of end-to-end/sequence-to-sequence models and neural vocoders enabled the realization of TTS that can produce speech that sounds as natural as human speech for well-articulated read speech [29, 30].

Non-Autoregressive Models

While autoregressive (AR) end-to-end/sequence-to-sequence models such as Tacotron 2 and Transformer TTS can produce very high quality speech, they take a relatively long time to produce speech because they predict a spectrum at a time step depending on the spectrum at the previous time step (autoregressive). To shorten synthesis time, several non-autoregressive (NAR) models have been proposed [77, 78, 79, 80, 81, 82]. The quality of speech produced by NAR models is also very high, but it is somewhat poorer than that of speech produced by AR models.

4.2 Why End-to-End/Sequence-to-Sequence TTS is Suitable for Synthesizing Rakugo Speech

We first attempted traditional pipeline/frame-by-frame neural speech synthesis models for synthesizing rakugo speech, but the quality of synthesized speech was very poor[1].

[1] Speech samples of pipeline/frame-by-frame models are available at https://nii-yamagishilab.github.io/samples-rakugo/pipeline/

In this thesis, we therefore chose an end-to-end/sequence-to-sequence model as an architecture of rakugo speech synthesis. There are two strong reasons underpinning the use of end-to-end models.

The first reason is that automatic phoneme segmentation and automatic f_o extraction are difficult for rakugo speech, especially for highly expressive speech. Pipeline/frame-by-frame models normally require phoneme boundary information as inputs [63]. But in the case of rakugo speech, the result of automatic estimation was extremely poor. In addition, f_o extraction also failed for non-speech sounds given by the performer, such as coughs, yawns, snores, sighs, and knocks, which sometimes play an important role in rakugo performance.

The second reason is that rakugo speech uses slightly old-fashioned Japanese dialects as mentioned in 2.2. In the case of pipeline/frame-by-frame TTS (Figure 4.4), an input text is first converted into linguistic features by the text analyzer. In the case of Japanese, the linguistic features are typically phonemes, accentual information, and other linguistic information, such as positional information, phrasal information, and parts of speech [83]. Japanese text analyzers process an input text based on grammatical rules and dictionaries. We need text analyzers for slightly old-fashioned Japanese dialects; Japanese text analyzers, however, are ordinary designed for processing modern standard Japanese. An example of incorrectly analyzed rakugo text is listed in Table 4.1.

On the other hand, end-to-end/sequence-to-sequence TTS does not have any explicit text processing modules. This means that it reduces the requirement for prepared linguistic information to synthesize speech. Namely, we do not need to prepare detailed information, i.e., accentual information[2], positional information, phrasal information, and parts of speech. This ease is very useful to rakugo speech synthesis because automatic text analysis is practically impossible for the Japanese language used in standards. The f_o extraction can also be avoided if the model predicts a mel spectrogram or waveform. Therefore, end-to-end/sequence-to-sequence TTS models are more suitable for rakugo speech.

[2] Many Japanese TTS studies using end-to-end/sequence-to-sequence models actually use accentual information as an input. This is because we can get more natural speech with input accentual information especially in terms of prosody. In this thesis, we do not use any accentual information to avoid time-consuming manual pitch accent annotation.

Table 4.1: Example of incorrectly analyzed rakugo text. Pitch falls at "＼," does not falls at "￣." "・" represents accentual phrase boundary.

Text	えーっと、おじいさんが、雪の降る山道をエッチラオッチラ歩ってると、オー大きな、真っ白な、首の長い鳥が、罠にかかって、バタバタバタバタ苦しんでいる。
Analysis result	エーット￣、オジ＼ーサンガ、ユキ＼ノ・フ＼ル・サンドーオ￣・エ＼ッチラオッチラ・フッテ＼ルト、オ＼ー・オ＼ーキナ、マッシ＼ロナ、クビノナガ＼イ・トリガ￣、ワ＼ナニ・カカ＼ッテ、バ＼タバタバタ・バ＼タ・クルシ＼ンデイル。
Expected result	エーット￣、オジ＼ーサンガ、ユキノ￣・フ＼ル・ヤマ＼ミチオ・エ＼ッチラオッチラ・アル＼ッテルト、オー￣・オ＼ーキナ、マッシ＼ロナ、クビノナガ＼イ・トリガ￣、ワ＼ナニ・カカ＼ッテ、バ＼タバタ・バ＼タバタ・クルシ＼ンデイル。

4.3 Relationship Between Speech Synthesis and Entertainment

4.3.1 Overview

As mentioned in 1.1, conventional speech synthesis studies do not focus on stirring listeners' emotions as a role of speech. However, singing voice synthesis, which is the process in which a machine synthesizes a singing voice and is a similar technology to speech synthesis, has already stirred listeners' emotions. Since the launch of Hatsune Miku [84], the most famous commercial singing voice synthesizer, in 2007, millions of performances using Hatsune Miku have been created, and singing voice synthesis has established its status as a culture today. She performed in tens of concerts in Japan in 2019 alone, and has been performed in concerts all over the world since 2014 [85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97]. In 2019, an attempt to revive a dead star singer voice succeeded, greatly impressing audiences [98, 99, 100]. The revived singer performed in the largest TV concert in Japan in 2019 [101].

On the other hand, in speech synthesis, there are no such entertaining performances at the professional level as far as we know. Speech synthesized entertaining performances including those of rakugo using TTS or singing voice synthesizer with many manual interventions are sometimes submitted to online video platforms [46, 47, 48], but unfortunately, in our opinion, the quality of them is far poorer than that of human professionals. We believe that speech synthesis should also have an ability to entertain audiences, and realization of this will enrich and change our lives and cultures.

4.3.2 Difference Between Rakugo and Audiobook Speech/TTS

Some readers may wonder how rakugo speech and its TTS differ from the speech of audiobooks, which is an active research topic in the speech synthesis field, and its TTS. An audiobook is a recording of a book read aloud. Recorded speech is conventionally human speech, but today synthesized speech is getting to be used as audiobook speech [20]. Audiobook speech synthesis has frequently been the theme of speech synthesis competitions [102, 103, 104, 105, 106, 107].

The main difference between rakugo speech synthesis and audiobook speech synthesis is that almost all parts of a rakugo story consist of conversations and dialogues between characters that are played by a performer from memory, and the conversations and dialogues make the story progress. As mentioned above, there are few narrative sentences in the conversational parts. In other words, rakugo performers should communicate the story to audiences without explicit explanations.

It should also be noted that rakugo speech is more casually pronounced than that of audiobooks because rakugo is performed from memory. In addition, expression in rakugo speech is far more diverse than that of audiobooks because the entire story is mostly comprised only of conversations and dialogues between characters. Also, as mentioned above, the Japanese language used in traditional rakugo stories is somewhat old-fashioned, and each character speaks a different dialect, sociolect, and idiolect according to his or her gender, age, social rank, and individuality.

Moreover, since rakugo is a form of entertainment, we argue that whether audiences are being entertained by rakugo speech or its TTS is essentially important. In addition, since the rakugo performance mainly consists of conversations and dialogues, building an appropriate rakugo TTS would also help us design improved conversational speech synthesizers.

5

Initial Modeling of Rakugo Speech Using Sequence-to-Sequence Speech Synthesis

5.1 Motivation

One of the main topics of this thesis is building rakugo speech synthesis. As described in 4.2, we first attempted traditional pipeline/frame-by-frame neural speech synthesis models to synthesize rakugo speech, but the synthesized speech had very poor quality; therefore we decided to use an end-to-end/sequence-to-sequence model as an architecture. On the other hand, we doubted that ordinary end-to-end/sequence-to-sequence model with a soft attention mechanism like Tacotron 2 could model rakugo speech well because rakugo speech is far more expressive than read-aloud speech.

We therefore attempted a new encoder-decoder sequence-to-sequence TTS without soft attention mechanism, which maps the encoder-decoder time steps but occasionally causes unacceptable errors because of poor training of the alignment, to deal with the casualness and diversity of rakugo speech. The content of this chapter is based on

[Pub1].

5.2 Models

5.2.1 SSNT-Based Speech Synthesis

In this section, we introduce the segment-to-segment neural transduction (SSNT) based TTS [108]. The SSNT-based TTS is an encoder-decoder model that can input a text or phoneme sequence and output a variable length of a mel spectrogram, but unlike all other encoder-decoder TTS [29, 30, 31, 33, 34, 109], the SSNT-based TTS does not use the soft attention mechanism, such as additive attention [110], forward attention [111], Gaussian mixture model (GMM) attention [112], location-sensitive attention [113], and self-attention [114]. This is because soft attention mechanisms are too flexible. Context vectors of an attention network are allowed to use input information at any time step of the encoder network. If the attention network is not well-trained, this results in unacceptable errors such as skipping input words, repeating the same phrases, and prolonging the same sounds.

Since a speech synthesis database normally has speech data and corresponding text aligned well, it is reasonable to have more explicit constraints. The SSNT-based TTS uses such explicit constraints. In SSNT [115], originally proposed for abstractive sentence summarization and morphological inflection in natural language processing, the decoder is allowed to consider the following two alignment options only: 1) Stay at the same encoder time step and increment the decoder time step and 2) transit to the next encoder time step and increments the decoder time step. It then computes a joint probability of an output feature sequence and the left-to-right self-transition alignment based on a trellis structure (Figure 5.1). The above motivation is very similar to the hidden Markov model (HMM), but SSNT uses neural networks to compute alignment probabilities in a nonlinear way.

The overall network structure of the SSNT-based TTS used in our experiments is shown in Figure 5.2. The main network is composed of an encoder and a decoder.

The structure of the encoder is the same as that of Tacotron 2 [29]. In the encoder, each input phoneme is embedded into a 512-dimensional vector. This vector can be concatenated to a style embedding, which will be described in 5.2.2, or a context

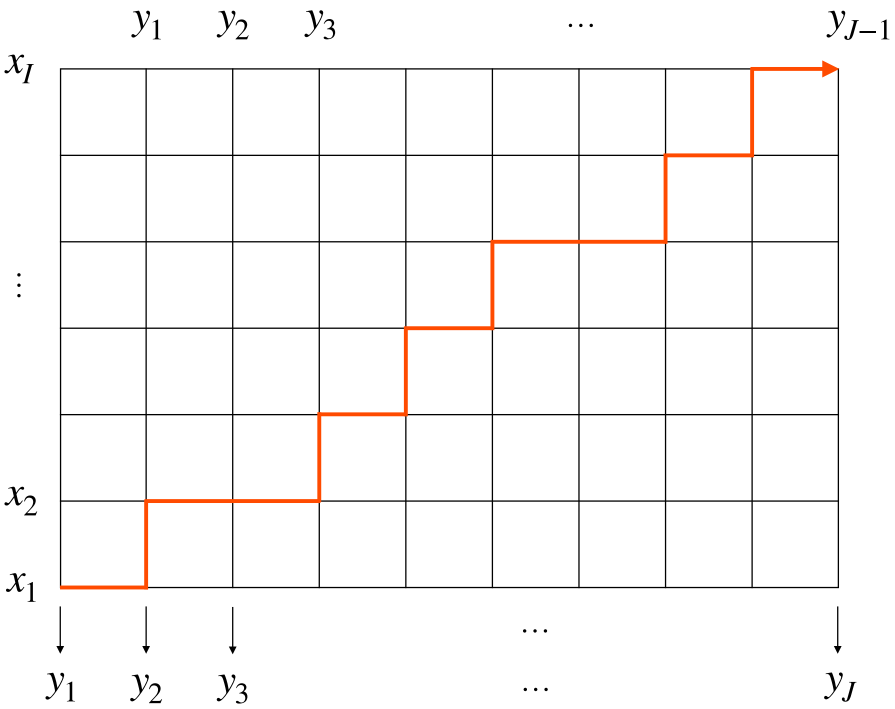

Figure 5.1: Trellis structure of SSNT-based TTS. x_i and y_j are input and output, respectively. Red path represents alignment between input and output.

embedding, which is derived from the input context labels listed in Table 3.3. The (concatenated) vector is input to three convolutional neural networks (CNNs) each containing 512 filters with a 5 × 1 shape (same padding [116]), followed by batch normalization [117] and rectifier linear units (ReLU) activations. The final output of the final CNN is then passed into a bi-directional [118] long short-term memory (LSTM) [70] that has 512 units (256 units per direction) to generate the final encoded features.

The output of the encoder is expected to have non-linearly encoded linguistic information. This is passed onto the decoder and concatenated with the information obtained from auto-regressive feedback. The predicted acoustic feature at the previous time step is fed back to a pre-net, a feed-forward network (FFN) that has 2 fully-connected layers of 256 ReLU units. This acts as an information bottleneck. The pre-net output is processed with 2 unidirectional LSTMs each containing 1,024 units. These LSTMs further consider contextual information of the feedback, and the output of the final LSTM is concatenated with the output of the encoder. This vector is further

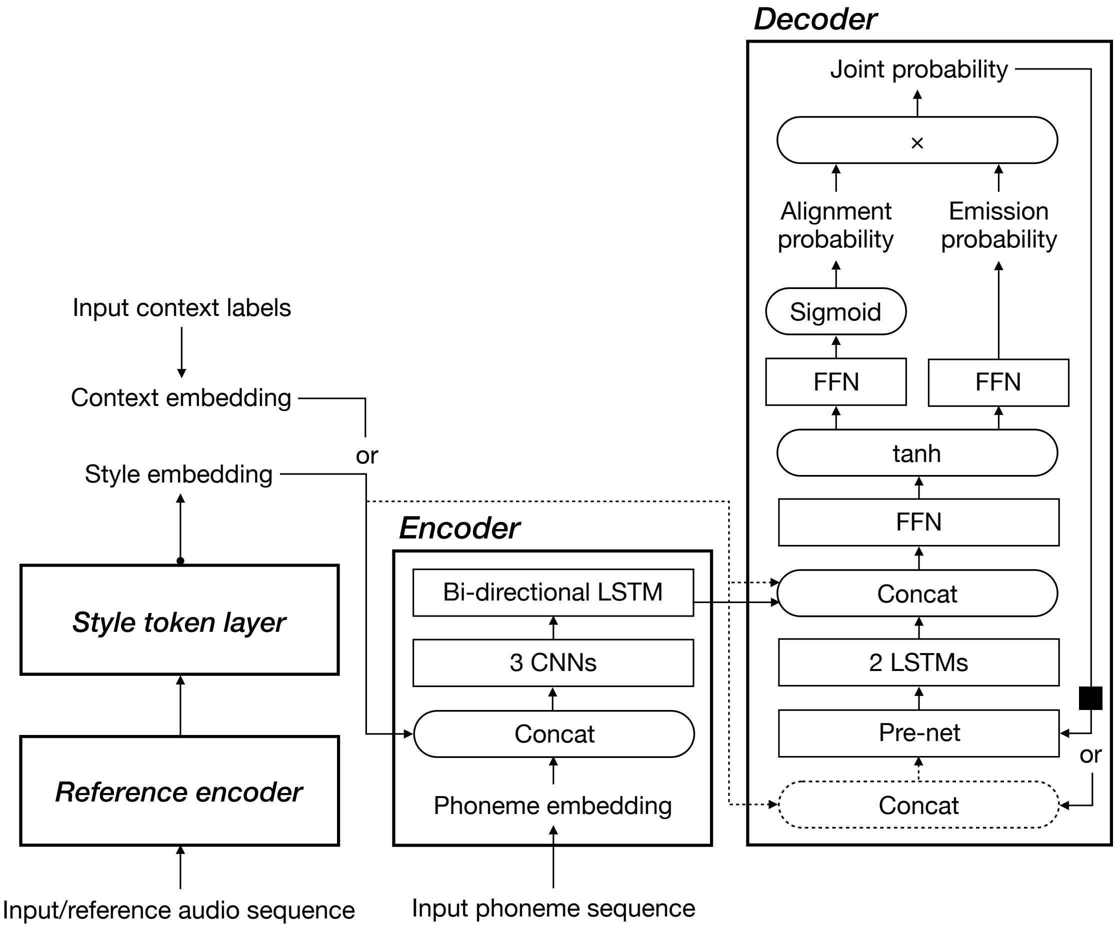

Figure 5.2: The overall network structure of the SSNT-based TTS. Dashed lines refer to SSNT-ATTR+ and SSNT-context+. ■ represents delay of one time step. Network structure of reference encoder and style token layer is shown in Figure 5.3.

transformed via an FFN that has 2 fully-connected layers of 256 tanh units, and is used as the basis of alignment and of emission networks. The alignment network predicts the above two transition probabilities via a sigmoid layer, and the emission network predicts a mel spectrum at the current decoder time step using an FFN that has 2 fully-connected layers of 80 linear units.

Training is conducted through minimizing the summation of the mean squared errors (MSEs) of the spectrograms. During training, dropout [119] is applied to each layer of the FFN in the pre-net and CNNs with probability 0.5 for regularization. Zoneout [120] is also applied to each LSTM with probability 0.1 for further regularization.

During the training, we used a forward-backward algorithm to optimize the network. During the synthesis, we used greedy decoding to generate speech. For more details, please refer to [108].

5.2.2 Global Style Tokens with SSNT-Based Speech Synthesis

We also used global style tokens (GSTs) [37] to enrich the speaking style of synthesized speech and make characters distinguishable from each other. The GST framework is a prosody transfer approach for end-to-end/sequence-to-sequence TTS. In this framework, we assume that TTS systems have access to a reference audio file from which we can borrow prosody and the speaking style, which are transferred to synthetic speech produced by the TTS system. Its role is to extract the prosody and speaking style that cannot be explained by the text input. The GST framework first extracts prosody from reference audio via a reference encoder and then creates a style embedding vector, which will be propagated to the decoder network. This can be easily integrated into the above SSNT-based TTS.

The architecture we used is basically the same as the original one, except for some parameters (Figure 5.3). An input or reference audio sequence, 80-dimensional mel spectrogram, is passed into a reference encoder. The reference encoder is composed of 6-layer 2D convolution layers with batch normalization and a 128-unit gated recurrent unit (GRU) [121]. Each convolution layer is made up of 3×3 filters with 2×2 stride (same padding) and ReLU activation. Batch normalization is applied to each layer. The number of filters in the layers are 128, 128, 256, 256, 512, and 512. The output of the final layer is passed into the GRU. The final state of the GRU is then passed into a network called the style token layer.

The style token layer is composed of 10 randomly initialized 512-dimensional embeddings called style tokens and a multi-head attention. The output of the reference encoder and tanh-activated tokens are then mapped by the multi-head attention. Any number of heads of the attention can be used if the dimension of style tokens can be divided by this number. If the number of heads is h, the dimension of tokens is $512/h$. The attention calculates the weights over tokens, and the weighted summation of tokens is treated as a style embedding, which will be concatenated to the phoneme embeddings output from the encoder of the SSNT-based TTS. The style embedding

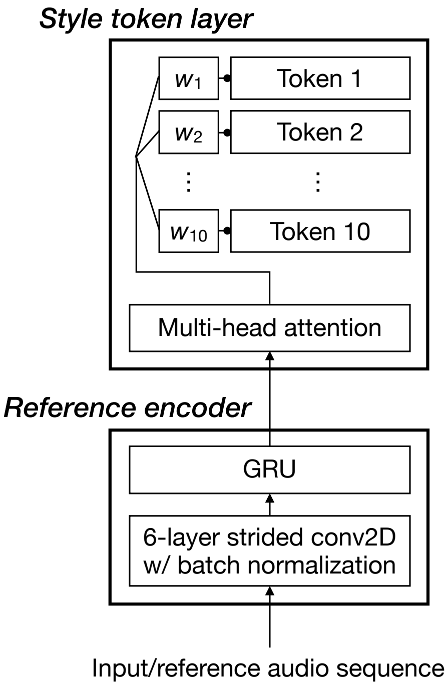

Figure 5.3: Network structure of reference encoder and style token layer.

vector is constant within a sentence.

5.2.3 Tacotron-Based Speech Sythesis

We used Tacotron-based models as references. The model architecture, shown in Figure 5.4, is based on Tacotron 2 [29] but a somewhat different. The main network is composed of an encoder, a decoder, and an attention network, which maps each time step of the encoder and that of the decoder. The encoder is the same as those of the SSNT-based model and Tacotron 2. The GST framework is not used in this model.

The output sequence of the encoder is used by an attention network that compresses the full encoded sequence as a fixed-length context vector for each decoder time step. We used forward attention with transition agent [111] instead of location sensitive attention [113], which is used in the original Tacotron 2, to learn the alignment between the encoder and decoder time steps more robustly and faster. The forward attention algorithm has a left-to-right initial alignment, which is useful because all the alignments of encoder-decoder TTS should be left-to-right because the output speech

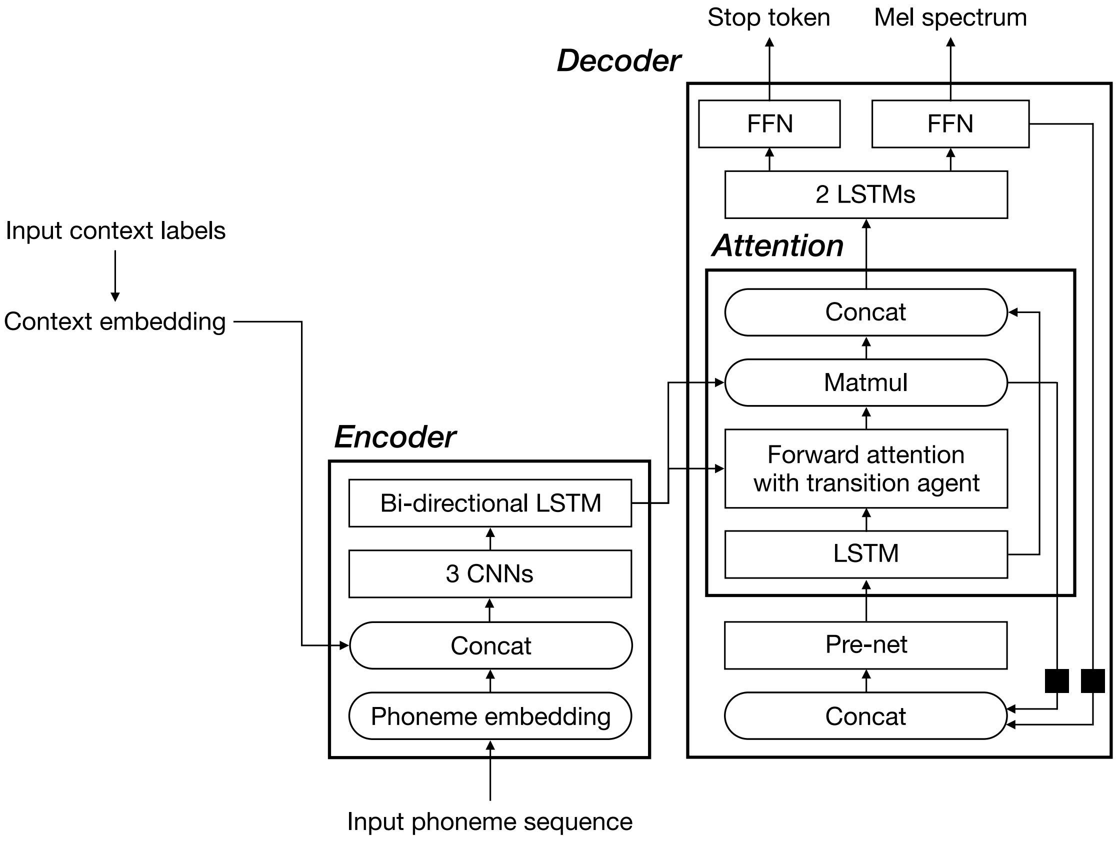

Figure 5.4: Overall network structure of Tacotron-based TTS. ■ represents delay of one time step.

has to be produced from the beginning to the end of the input text. For more details, please refer to [111].

In the decoder, the predicted mel spectrum in the previous time step is first passed into a pre-net, an FFN that has 2 fully-connected layers of 256 ReLU units. The pre-net output and the context vector from the attention network at the previous time step are concatenated and passed into 2 unidirectional LSTMs each containing 1,024 units. Using the output sequence of the LSTMs and encoder output, the context vector at the current time step is calculated. Then the concatenation of the output of LSTMs and context vector is passed into an FFN that has a fully-connected layer of 80 linear units to generate a mel spectrum. A post-net is not used unlike the original Tacotron 2 to match to the SSNT-based model, which does not have a post-net.

In parallel with the spectrogram prediction, the concatenation of the output of decoder LSTMs and the attention vector is projected to a scalar and activated by a sigmoid function to predict the probability of the completion of the output sequence. This probability is called the "stop token."

Training is conducted through minimizing the summation of the MSEs of the spectrograms, the binary cross entropy loss of the stop token, and the L2 regularization loss. During training, dropout is applied to each layer of the FFN in the pre-net with probability 0.5 for regularization. Zoneout is also applied to each LSTM with probability 0.1 for further regularization.

5.3 Experiments

5.3.1 Purpose of Experiments

We modeled rakugo speech with two different types of models, SSNT-based (5.2.1) and Tacotron-based (5.2.3) models. We first analyzed alignment errors of synthesized speech. Almost all the encoder-decoder sequence-to-sequence TTS including the Tacotron-based one above use a soft attention mechanism to map each time step of the encoder and that of the decoder. In speech synthesis, alignments should be left-to-right. The soft attention mechanism does not have such a restriction, so it may cause alignment errors. The stop token used in the Tacotron-based TTS may fail to predict the termination of a sentence. The SSNT-based TTS does not incur any alignment error after infinite iterations of training. However, because we can actually train models in finite iterations, it may incur alignment errors of incompleteness.

We also conducted a listening test to compare their performances and assess how they are accepted by the public. Since a rakugo paragraph is composed of conversations or dialogues between characters, we are interested in not only the naturalness of synthesized speech, but also how accurately listeners distinguish characters, how well listeners understand the content of the speech.

5.3.2 Models and Samples Used in the Experiments

We used 16 fully-annotated stories out of the total 25 stories in the Database I[1]. The 16 stories are about 4.31 hours long in total, except for pauses between sentences, and contain 7,337 sentences. We used 6,459 sentences for training (3.74 hours), 717 for validation (0.42 hours), and 161 for testing (0.15 hours). The training and validation sets did not include very short (< 0.5 s) or very long (≥ 20 s) speech to reduce alignment errors during training. It should be noted that the total amount of speech was rather small. We attempted to build Tacotron-based models with a larger amount of speech containing the very short or very long speech above, and we also attempted to fine-tune from well-trained Tacotron-based models trained with read speech [122], but their quality was similar to or worse than those that produced the results below.

We trained several models for our experiments.

- **SSNT** is an SSNT-based model, and no style embeddings or context features are input.
- **SSNT-GST-*n*** is an SSNT-based model with GSTs with an *n*-head multi-head attention. We used 4, 8, 16, 32, and 64 as *n*. It should be noted that the reference audio of the test sets is natural speech itself[2].
- **SSNT-ATTR** is an SSNT-based model with manually labeled context features belonging to ATTR (role and individuality) only. The dimension of the context embedding is 4.
- **SSNT-context** is an SSNT-based model with all the manually labeled context features. The dimension of the context embedding is 68.
- **SSNT-ATTR+** and **SSNT-context+** are the same models as SSNT-ATTR and SSNT-context, respectively, except for the additional concatenation of the context embedding with encoder outputs and feedback to the decoder (dashed lines shown in Figure 5.2).

[1] Annotation of the database is a work in progress

[2] This is a choice by design since this makes comparisons of *-GST-*n* with other systems that use manually labeled contest features fairer.

- **Tacotron**, **Tacotron-ATTR**, and **Tacotron-context** are the same models as SSNT, SSNT-ATTR, and SSNT-context, respectively, but the Tacotron-based model is used instead of SSNT.

We trained each model with about 3,500 epochs (mini-batch size: 32, 700,000 steps (SSNT), mini-batch size: 96, 250,000 steps (Tacotron)) with an initial learning rate of 0.0001, and the learning rate was decayed exponentially. The optimization method was Adam. The number of mel filters for spectrograms was 80. The spectrograms were converted from 48 kHz/16 bit waveforms with 50-ms-long frame, 12.5-ms frame shift, Hann window, and 4,096-long fast Fourier transform. The waveforms were normalized to −26 dBov by sv56 [123] in each sentence. Values of the spectrograms were transformed into 0 mean and 1 standard deviation at each dimension over all the data.

Predicted mel spectrograms were converted into waveforms by using a WaveNet vocoder trained with natural mel spectrograms and waveforms of all the training, validation, and test sets. The sampling rate of the waveform was 16 kHz[3].

5.3.3 Alignment Error Analysis

Details of Analysis

As objective evaluation, we first compared the rates of obvious alignment errors of the SSNT-based and Tacotron-based systems. We used the test sentences and first generated an alignment probability of each of the input phoneme sequences. We then computed the expected values of the alignment probability over decoder time steps. This tells us which phoneme input within a sentence is used at each encoder time step in general. Because the phoneme sequence should be used from the beginning to the end sequentially in order, we can easily tell that an obvious error happens if adjacent encoder time steps have a large gap (e.g. four time steps or more) or if the next encoder time step is not a monotonic increase (e.g. minus two time steps). Moreover, we counted it as an error if the last encoder time step does not correspond to the last decoder time step. It may incur in the SSNT-based systems because we can train them

[3] Speech samples are available at https://nii-yamagishilab.github.io/samples-rakugo/201909_SSW10/.

Table 5.1: Alignment error rates (%) for the test set.

System	
SSNT	1.86
SSNT-GST-4	2.48
SSNT-GST-8	1.86
SSNT-GST-16	3.11
SSNT-GST-32	1.86
SSNT-GST-64	3.73
SSNT-ATTR	2.48
SSNT-ATTR+	**1.24**
SSNT-context	1.86
SSNT-context+	**1.24**
Tacotron	28.57
Tacotron-ATTR	29.19
Tacotron-context	26.09

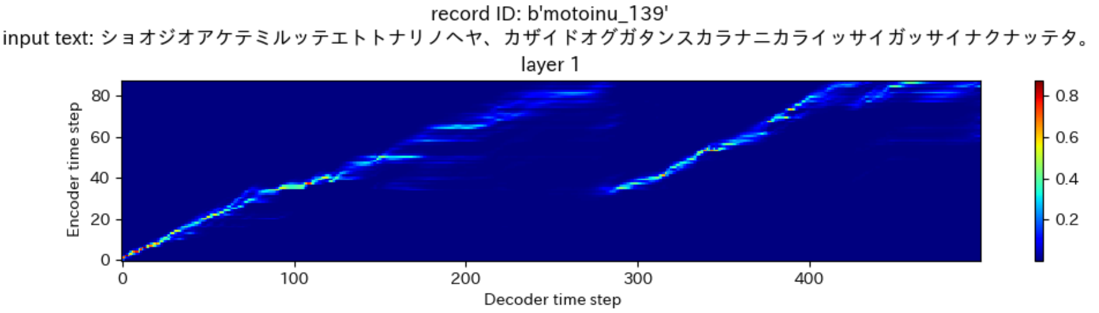

Figure 5.5: Example of obvious alignment errors on a Tacotron-based model.

in finite iterations though they always reach the last decoder time steps with infinite iterations of training.

Results

Alignment error rates per system are listed in Table 5.1. We can see that the SSNT-based systems greatly reduced the alignment errors. SSNT-ATTR+ and SSNT-context+ got the lowest error rates. Additional concatenation of context embeddings with encoder outputs and feedback to the decoder seems to reduce alignment errors further.

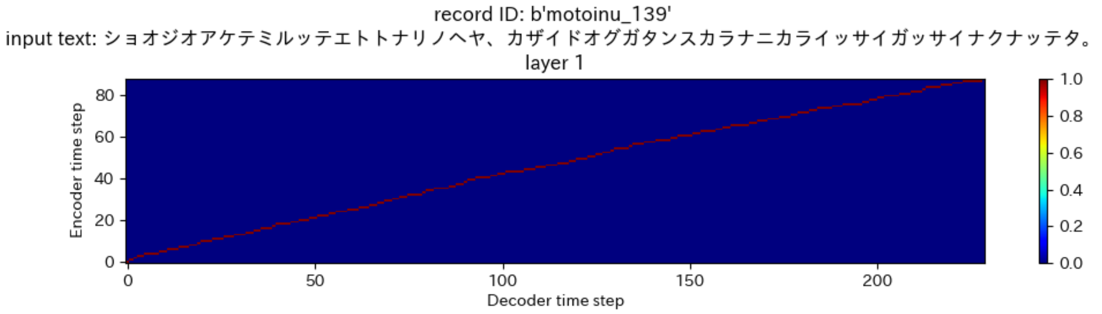

Figure 5.6: Alignment of an SSNT-based model for the same sentence as 5.5

Discussions

The SSNT-based models had very low alignment error rates. We think this is because alignment transition in SSNT is more restricted than that of the soft attention mechanism. The additional concatenation probably helped the alignment network to use context features well. Figure 5.5 is an example of obvious alignment errors caused by a Tacotron-based model. In contrast, the SSNT-based models aligned the encoder-decoder time steps well (Figure 5.6).

5.3.4 Listening Test

Test Conditions

We selected a set of sentences comprising a short story as materials for the listening test. A total of 161 sentences consisting of 12 stories[4] were prepared, and sentence-level audio files were concatenated as one audio file per story. Because the speech samples were predicted sentence by sentence, and pauses between sentences were not predicted, the pauses between sentences used in the test were the same as those of real audio recordings. These should be predicted by systems, but that is out of the scope of this thesis. Listeners evaluated speech NOT in sentence-by-sentence samples but in a whole story. Natural speech recordings are not included in our test because we wanted to see the differences between the systems more precisely instead of the differences between TTS and natural speech.

[4] The details of the test stories are described in Appendix A.

We conducted a five-scale mean opinion score (MOS) test. In each evaluation round, listeners listened to the same short story generated using one of the models listed in 5.3.2 in each screen. For each round, the permutation of the systems were randomly selected. One of the speech was presented on each screen, and listeners answered three MOS-based questions:

1) How natural did the speech sound? (naturalness)

2) How accurately did you think you could distinguish each character?

3) How well did you think you could understand the content?

A total of 135 listeners conducted 453 evaluation rounds through crowdsourcing.

Results

The results are shown in Figure 5.7.

For statistical analysis, we conducted Brunner-Munzel tests [124] with Bonferroni correction among the scores for all the model combinations. We can see that the proposed SSNT-based models had higher scores than the Tacotron-based models in all three questions. Among the SSNT-based models, SSNT-GST-8 slightly outperformed the others. GSTs generally could increase the quality of speech and its representation, but too many heads of multi-head attention seem to reduce quality. The same can be said about manually labeled context features (SSNT-ATTR(+) vs. SSNT-context(+)). Additional concatenation of the context embeddings with encoder outputs and feedback to the decoder did not increase the scores (e.g. SSNT-ATTR vs. SSNT-ATTR+) though it reduced alignment errors.

Discussions

The SSNT-based models outperformed the Tacotron-based models. We think this is obviously because the alignment error rates for the SSNT-based models is far lower than those for the Tacotron-based models as mentioned in 5.3.3. The reduction of scores with too many heads of multi-head attention in SSNT-GST-*n* was probably caused by overfitting. Context features did not increase the quality, and too many contexts may also drop the quality as SSNT-context and SSNT-context+ did, also probably because of overfitting.

The scores of question 2 (distinguishability of characters) and question 3 (understandability of content) showed similar trends. A better distinction of each character and a better understanding of the content are probably correlated with each other.

5.4 Summary

We could successfully build initial rakugo speech synthesis using the SSNT-based TTS. The SSNT-based models used in the experiments could learn the encoder-decoder alignment well, and produced somewhat natural, character-distinguishable, content-understandable speech. However, the MOSs of the best model, SSNT-GST-8, were just around 3.

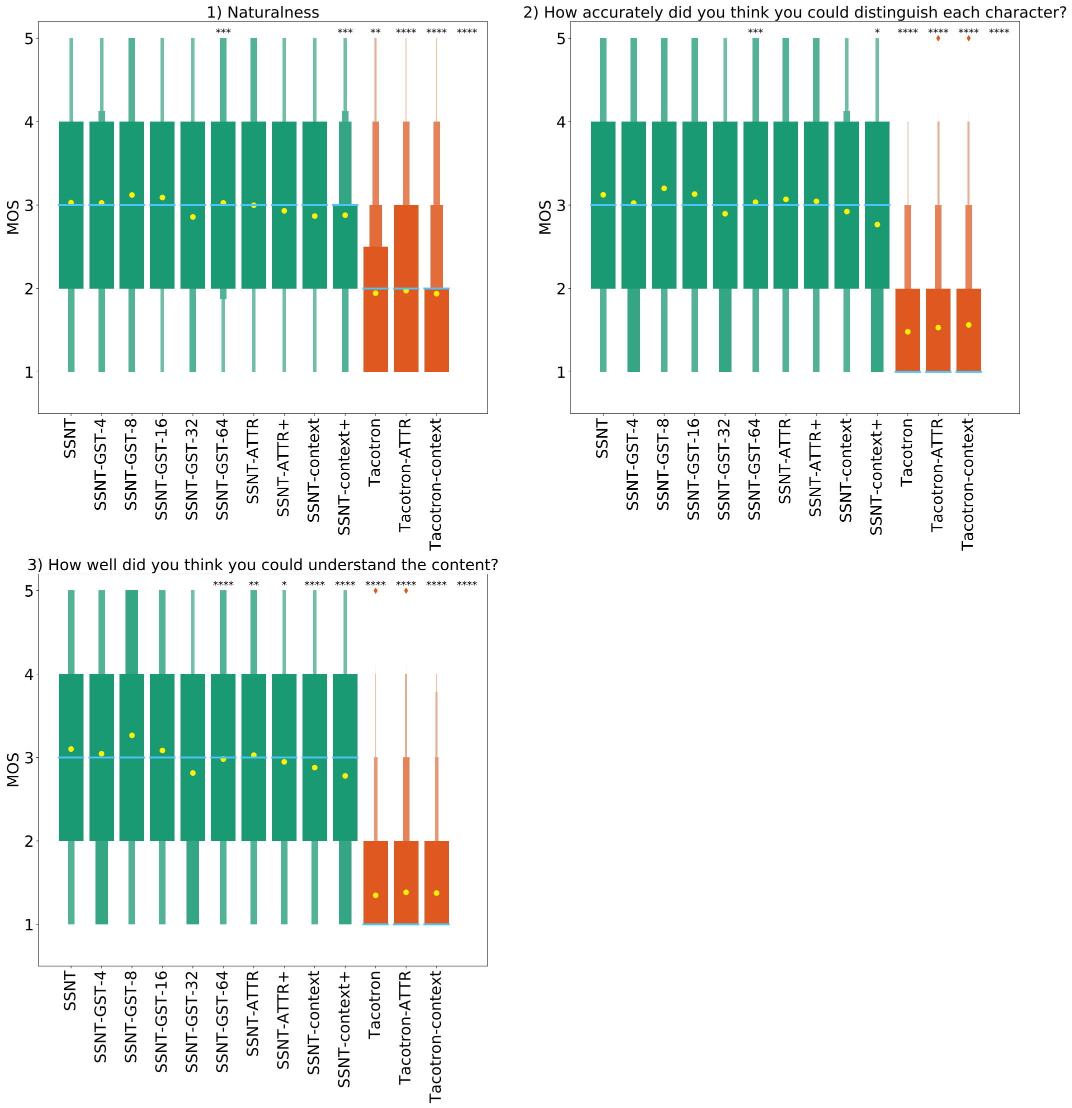

Figure 5.7: Boxen plots for each question of listening test. Light blue lines and yellow dots represent medians and means, respectively. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$. Only significant differences via Brunner-Munzel test with Bonferroni correction between SSNT-GST-8 and each system are shown.

6

Modeling of Rakugo Speech Using Tacotron 2 with Self-Attention, Global Style Tokens, and Manually Labeled Context Features

6.1 Motivation

After the submission of [Pub1], on which the contents of Chapter 5 are based, we refined[1] our implementation of Tacotron [125, 126], and Tacotron became able to learn the encoder-decoder alignment well and produce speech of better quality than that of

[1] The detail of the refinement is shown in Figure 6.1.

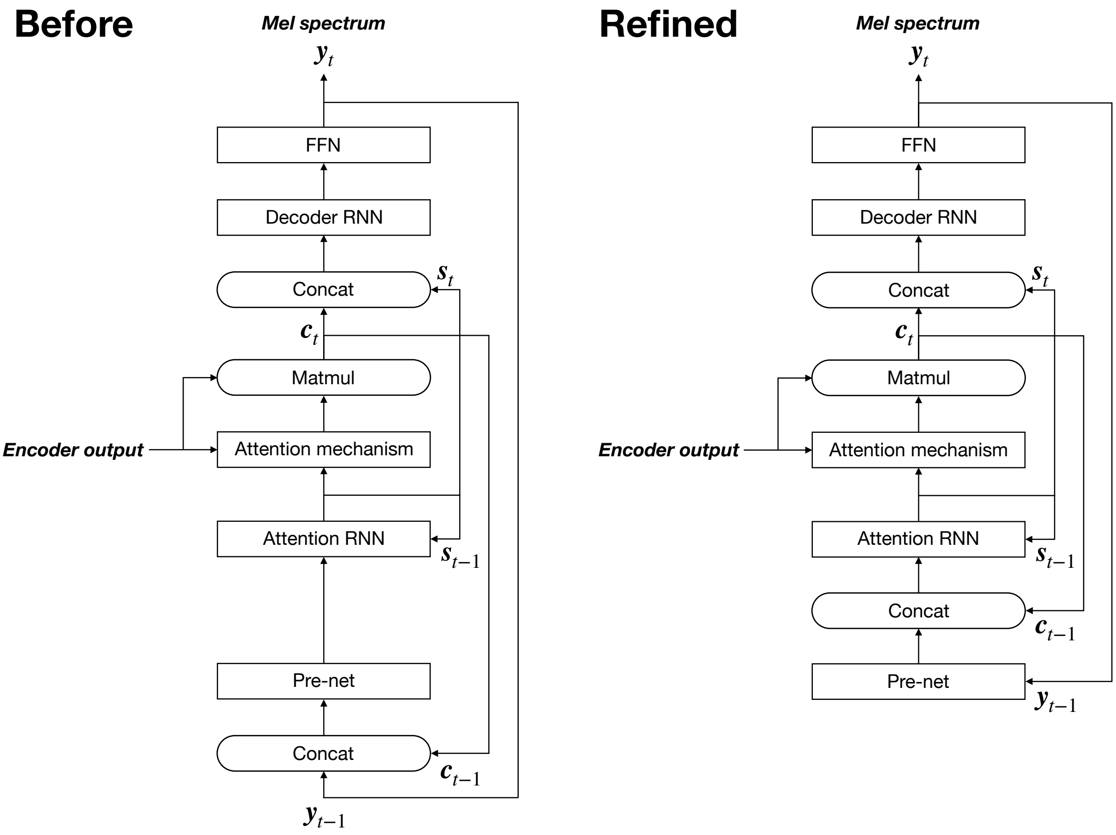

Figure 6.1: Our implementation of Tacotron before/after refinement. $\boldsymbol{y}_t$: output mel spectrum at time t, $\boldsymbol{c}_t$: context vector of attention at time t, $\boldsymbol{s}_t$: output of attention RNN at time t. The key is that dropout is applied in the pre-net during training. The implementation before the refinement (left), $\boldsymbol{y}_{t-1}$ and $\boldsymbol{c}_{t-1}$ are concatenated and then input to the pre-net; therefore the information $\boldsymbol{c}_{t-1}$ has will be regularized and partly forgot because of the dropout during training.

the SSNT-based TTS[2] (Figure 6.2). In this chapter, we attempt the (refined) Tacotron and an additional architecture to further improve quality of synthesized rakugo speech.

In Chapter 5, we successfully built initial rakugo speech synthesis system which produced somewhat natural, character-distinguishable, and content-understandable speech, but unfortunately the quality of speech was not so fine. Also, the refinement of our implementation enabled Tacotron to synthesize speech of better quality than the

[2] It should be noted that the implementation before the refinement seemed well in the case of ordinary speech, as in [127]. The casualness and diversity of rakugo speech certainly make it difficult to build TTS models properly.

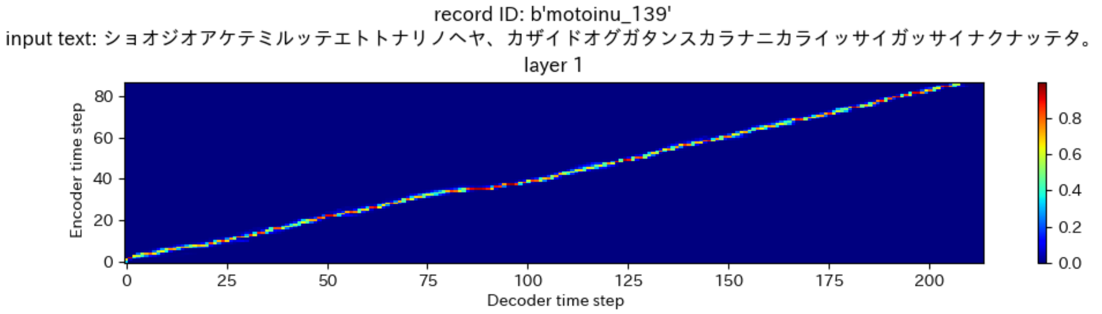

Figure 6.2: Alignment of the refined Tacotron for the same sentence as 5.5

SSNT-based TTS.

In this chapter, we attempt Tacotron 2 [29], a state-of-the-art speech synthesis system that produces read-aloud speech that sounds as natural as human speech, and an enhanced version with self-attention [114] to further improve the quality of speech, inspired by [127]. The content of this chapter is based on [Pub2].

6.2 Models

6.2.1 Tacotron 2

As mentioned in 6.1, Tacotron 2 is a state-of-the-art speech synthesis system that produces read-aloud speech that sounds as natural as human speech. Some Tacotron-based systems can model expressive speech, including audiobook speech well [37, 38, 40, 128]. We therefore argue that modeling rakugo speech using Tacotron 2 is reasonable.

The architecture of the Tacotron-2-based rakugo TTS model is shown in Figure 6.3. The main network of this model is strictly the same as the original Tacotron 2 except that we used forward attention with transition agent instead of location sensitive attention to learn alignments more robustly and faster.

The main network is composed of an encoder, a decoder, and an attention network, which maps each time step of the encoder and that of the decoder. The detail of the encoder and attention network is described in 5.2.3.

The decoder architecture is strictly the same as that of the original Tacotron 2

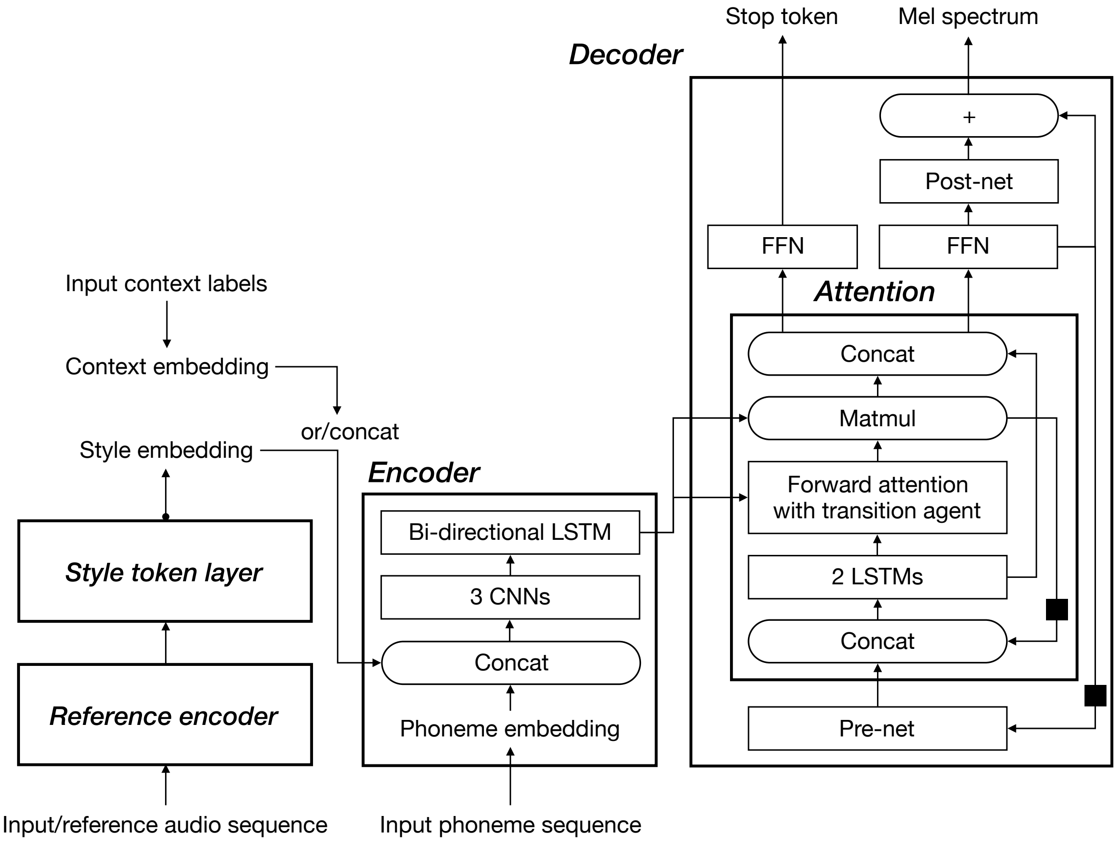

Figure 6.3: Overall network structure of Tacotron-2-based rakugo TTS. ■ represents delay of one time step. Network structure of reference encoder and style token layer is shown in Figure 5.3.

unlike the Tacotron-based TTS used in Chapter 5. In the decoder, the predicted mel spectrum in the previous time step is first passed into a pre-net, an FFN that has 2 fully-connected layers of 256 ReLU units. The pre-net output and the context vector from the attention network at the previous time step are concatenated and passed into 2 unidirectional LSTMs each containing 1,024 units. Using the output sequence of the LSTMs and the encoder output, the context vector at the current time step is calculated. Then the concatenation of the output of the LSTMs and the context vector is passed into an FFN that has a fully-connected layer of 80 linear units to generate a mel spectrum. To predict a residual for improving the reconstruction, the predicted spectrum is passed into a post-net, 5 CNNs each containing 512 filters with a 5×1 shape (same padding) followed by batch normalization. Then tanh activations are applied except for the final layer. The summation of the former predicted mel spectrum and the output of the post-net is the final output mel spectrum.

In parallel with the spectrogram prediction, the concatenation of the output of the decoder LSTMs and the attention vector is projected to a scalar and activated by a sigmoid function to predict the probability of the completion of the output sequence. This probability is called the “stop token.”

Training is conducted through minimizing the summation of the MSEs of the spectrograms both before and after the post-net, the binary cross entropy loss of the stop token, and the L2 regularization loss. During training, dropout is applied to each layer of the FFN in the pre-net and CNNs in the post-net with probability 0.5 for regularization. Zoneout is also applied to each LSTM with probability 0.1 for further regularization.

6.2.2 Tacotron 2 with Self-Attention

For further improvement, we enhanced the Tacotron 2 above with self-attention [114] (SA-Tacotron). Self-attention can effectively capture long-term dependencies, and it was reported that the enhanced version of Tacotron with self-attention exceeds the original Tacotron [33] for Japanese [127]. We therefore enhanced the Tacotron 2 above with self-attention. The network structure is shown in Figure 6.4.

The network structure is similar to that presented in [127], except that the encoder and decoder used in our study are based on Tacotron 2, while those in [127] are based on Tacotron.

In the encoder, a self-attention block is inserted after the bi-directional LSTM. A self-attention block consists of a self-attention, followed by a fully-connected layer with tanh activation and residual connections. This block is expected to capture the long-term dependency inside the input phoneme sequence. The number of heads in the multi-head attention [129] in the self-attention is 2, and the dimension is 32. During training, dropout is applied to the multi-head attention with probability 0.05 for regularization. The encoder generates two output sequences, one is from the bi-directional LSTM, and the other from the self-attention block.

The output sequences of the encoder are input into two attention networks. The output sequence from the bi-directional LSTM is used by a forward attention with transition agent, the same architecture as the Tacotron 2 described in 6.2.1. On the other hand, the output sequence from the self-attention block is used by an additive attention [110]. The context vectors from the two attentions are concatenated and used in the decoder.

The structure of the decoder is the same as that of the decoder of the Tacotron 2 described in 6.2.1, except that a self-attention block is inserted after the LSTMs. This block is expected to capture the long-term dependency inside the output sequence. The number of heads in the multi-head attention in the self-attention is 2, and the dimension is 1,568 (1024 + 512 + 32).

6.2.3 Global Style Tokens with Tacotron 2 and SA-Tacotron

We also use GSTs to enrich the speaking style of synthesized speech and make characters distinguishable from each other, as the case of the SSNT-based TTS in Chapter 5. The architecture is the same as that described in 5.2.2.

We use 8 heads in this chapter based on the results of preliminary experiments. The estimated style embedding will be concatenated to the phoneme embeddings output from the encoder of the Tacotron-2-based or SA-Tacotron-based model. The style embedding vector is constant within a sentence.

6.3 Listening Test

6.3.1 Purpose of Listening Test

We modeled rakugo speech with two different types of models, Tacotron-2-based (6.2.1) and SA-Tacotron-based (6.2.2) models. We conducted a listening test to compare their performances and assess how they are accepted by the public.

The same as Chapter 5, since the main part of a rakugo story is composed of the conversations or dialogues between the characters, we asked listeners not only the naturalness of synthesized speech, but also how accurately they distinguish characters, how well they understand the content of the speech. In this chapter, we also asked them how well the speech entertained them considering that rakugo is a form of entertainment.

6.3.2 Models and Samples Used in the Listening Tests

The same as Chapter 5, we used 16 fully-annotated stories out of the total of 25 stories in the Database I for our listening test. The 16 stories are about 4.31 hours long in total, except for pauses between sentences, and contain 7,341 sentences[3]. We used 6,437 sentences (3.74 hours) for training, 715 for validation (0.40 hours), and 189 (0.17 hours) for testing. The training and validation sets did not include very short (< 0.5 s) or very long (≥ 20 s) utterances to reduce alignment errors during training.

We trained several models for the listening test.

- **Tacotron** is a Tacotron-2-based model, and its input is the phoneme sequence. No style embeddings or context features were used as an input.
- **Tacotron-GST-8** is a Tacotron-2-based model including GSTs with an 8-head attention. It should be noted that the reference audio of the test set was natural speech[4].
- **Tacotron-ATTR** is a Tacotron-2-based model with manually labeled context features belonging to ATTR (role and individuality) only. The dimension of the context embedding is 4.
- **Tacotron-context** is a Tacotron-2-based model with all of the manually labeled context features. The dimension of the context embedding is 68.
- **Tacotron-GST-8-ATTR** and **Tacotron-GST-8-con- text** are Tacotron-2-based models with a combination of GSTs with an 8-head attention and manually labeled context features belonging to ATTR or all contexts, respectively.
- **SA-Tacotron**, **SA-Tacotron-GST-8**, **SA-Tacotron-ATTR**, **SA-Tacotron-context**, **SA-Tacotron-GST-8-ATTR**, and **SA-Tacotron-GST-8-context** are the same models as Tacotron, Tacotron-GST-8, Tacotron-ATTR, Tacotron-context, Tacotron-GST-8-ATTR, and Tacotron-GST-8-context, respectively, but they are based on SA-Tacotron instead of Tacotron 2.

[3] The number of sentences is different from that in Chapter 5 because we reconsidered the point of splitting sentences for some sentences.

[4] We used natural speech as the reference audio of the test set to compare *-GST-8 with other models (*-ATTR and *-context) that use manually labeled correct context features, which were labeled through listening to recorded speech of the test set, in a fair condition.

We trained each model for about 2,000 epochs. The mini-batch size was 128. The initial learning rate was 0.001, and the learning rate was decayed exponentially. The optimization method was Adam, and the number of mel filters for input spectrograms was 80. The spectrograms were generated from 48 kHz/16 bit waveforms with 50-ms-long frame, 12.5-ms frame shift, Hann window, and 4,096-long fast Fourier transform. The waveform were not normalized unlike Chapter 5 because the training results were fine without normalization. Values of the spectrograms were normalized into 0 mean and 1 standard deviation at each dimension over all the data.

Predicted mel spectrograms were converted into waveforms by using a WaveNet vocoder trained with natural mel spectrograms and waveforms of all the training, validation, and test sets[5]. The sampling rate of the output waveform was 16 kHz[6].

6.3.3 Test Conditions

We selected a set of sentences comprising short stories as materials for the listening test. A total of 189 sentences comprising 13 short stories[7] were prepared, and sentence-level audio files were concatenated as one audio file per story. The duration of the stories ranges from 11 seconds to 1 minute and 58 seconds (total 11 minutes and 14 seconds) in the case of real recordings. Because the speech samples were predicted sentence by sentence, and pauses between sentences were not predicted, the pauses between sentences used in the test were the same as those of real audio recordings. In other words, pauses between sentences were taken from real speech, and other prosody including intra-sentence pauses were predicted using models. It is obvious that pauses between sentences should also be predicted using models, but that is out of the scope of this paper. Listeners evaluated speech NOT sentence by sentence but in a whole story. Analysis-by-synthesis (AbS; copy synthesis) speech was also used for the test. The concatenated audio files were normalized to −26 dBov by sv56.

[5] We used natural mel spectrograms and waveforms of not only the training and validation sets but also the test set for the WaveNet training. This design is to make comparison with AbS speech fairer. It should be noted that we also compared the above WaveNet to one without the test set and made sure that there was no perceptual difference.

[6] The sampling rate of the output waveform was 16 kHz although that of the predicted mel spectrograms was 48 kHz. This is because of computational complexity of the WaveNet model. Speech samples are available at https://nii-yamagishilab.github.io/samples-rakugo/\201910_IEEE_access/

[7] The details of the test stories are described in Appendix A.

We used MOS as the metric for the test. In each evaluation round, listeners listened to the speech of all 13 stories each synthesized using one of the models listed in 6.3.2 or AbS speech. For each listener, the story-system combinations and their permutation were randomly selected. The audio of one story was presented on each screen, and listeners answered four MOS-based questions:

1) How natural did the speech sound? (naturalness)

2) How accurately did you think you could distinguish each character?

3) How well did you think you could understand the content?

4) How well were you entertained?

We used a five-point MOS scale. A listener was allowed to answer only one evaluation round because listeners would remember the content of the stories. A total of 183 listeners participated in 183 evaluation rounds through crowdsourcing.

6.3.4 Results

We should carefully compare the scores because listeners may evaluate the speech with more diverse values than the cases of ordinary listening tests for measuring naturalness. Considering this, for fair comparison, the scores were first normalized to 0 mean and 1 standard deviation for each listener to absorb variations of scores among listeners, and then further normalized per story so that the mean score of the AbS of human performance would be 0 and the standard deviation of it would be 1 to diminish the effects of the content of the story.

The results are shown in Figure 6.5. For statistical analysis, we conducted Brunner-Munzel test with Bonferroni correction among the (normalized) scores for all the model combinations. For Q1–Q3, AbS speech was superior to all the (SA-)Tacotron-based models. Regarding Q4, AbS speech was also superior to all these models, but the differences between some models and AbS speech were smaller. We could not find any significant differences among (SA-)Tacotron-based models.

We compared the results systematically for Tacotron and SA-Tacotron, with and without GST and/or context labels, but we found no significant trends.

6.3.5 Relationship Between Obvious Errors and the Scores

To investigate the relationship between obvious errors and the scores, we calculated alignment and pitch-accent error rates for the test set.

To calculate alignment error rates, the author carefully listened to all the synthesized speech for the test sentences produced with each model and checked whether any alignment errors occur per sentence. Prolonging phonemes, skipping phonemes, repeating phrases, and late terminations were defined as alignment errors.

The author also carefully listened to all the synthesized speech for the test sentences produced with each model and counted the number of pitch-accent errors. We regarded a pitch-accent phrase pronounced in a different accent from that of the recorded speech as an accent error even if the synthesized accent is acceptable as natural Japanese. The total number of pitch-accent phrases in the test set was 1,089.

Alignment and pitch-accent error rates for the test set are listed in Tables 6.1 and 6.2, respectively. We can see that both error rates differed by system. However, there were no obvious relationships between alignment or pitch-accent error rates and the MOSs of the listening test.

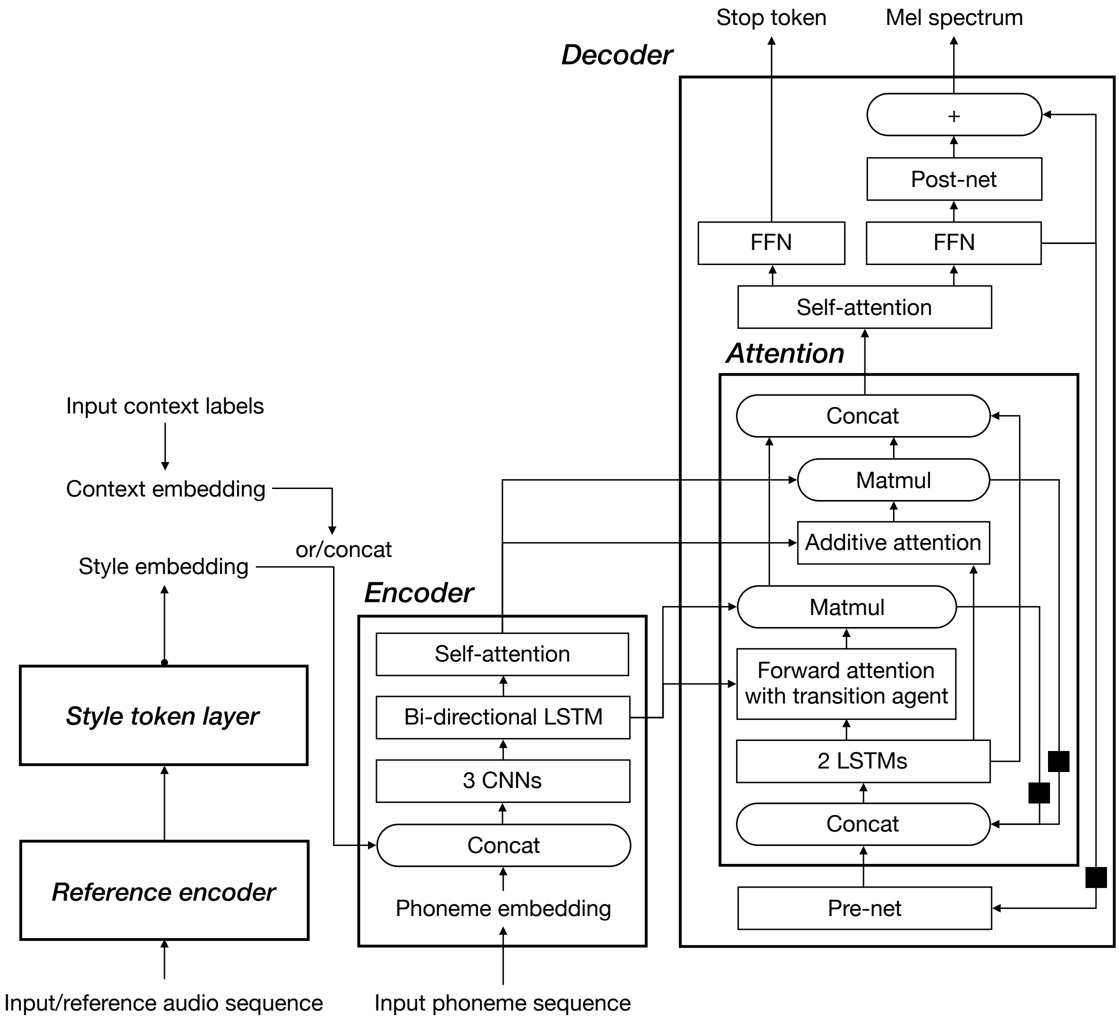

Figure 6.4: Overall network structure of SA-Tacotron-based rakugo TTS. ■ represents delay of one time step. Network structure of reference encoder and style token layer is shown in Figure 5.3.

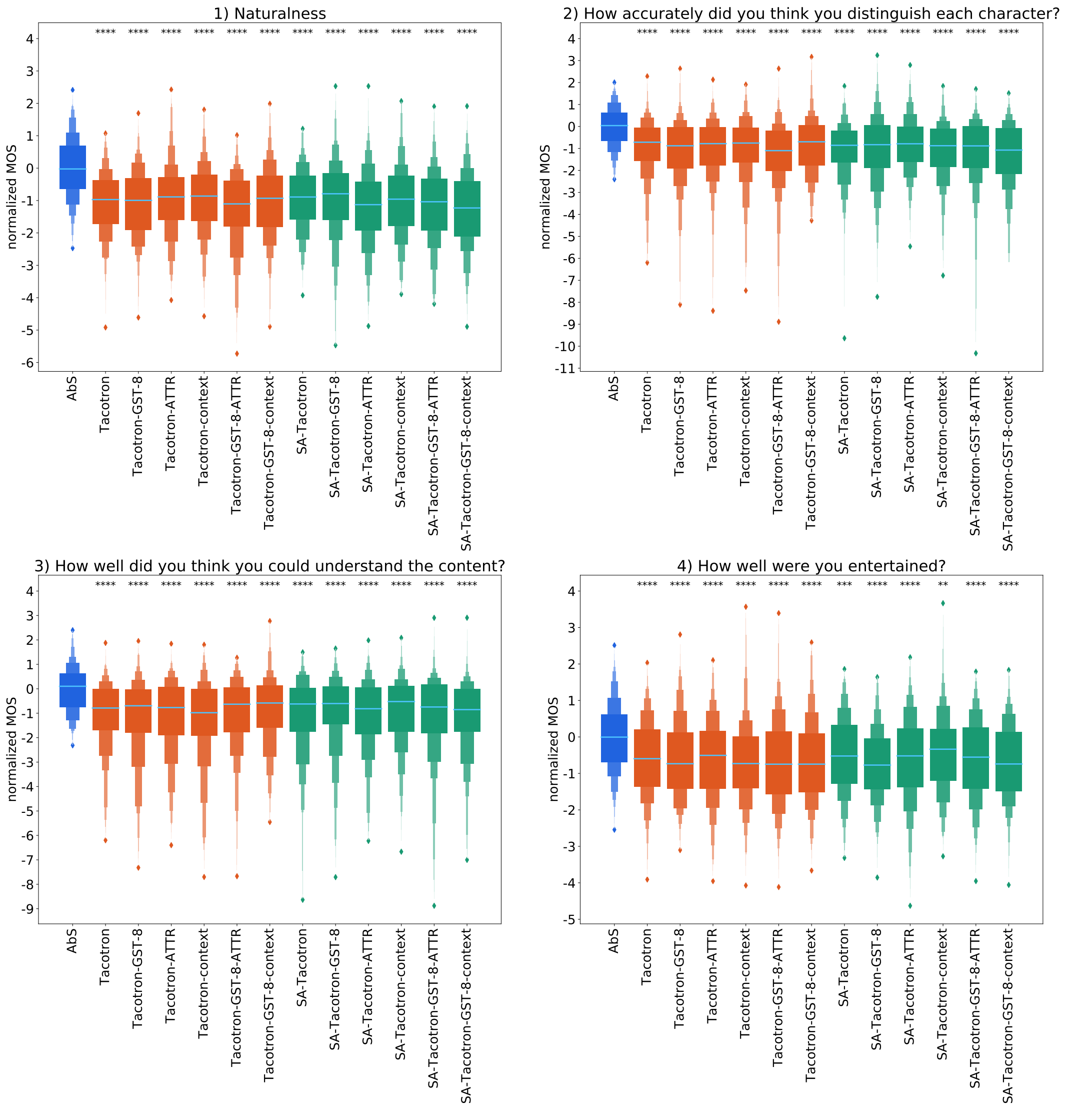

Figure 6.5: Boxen plots for each question of listening test. Light blue lines represent medians. **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$. Only significant differences via Brunner-Munzel test with Bonferroni correction between AbS speech and each (SA-)Tacotron system are shown. There were no significant differences among (SA-)Tacotron-based models.

Table 6.1: Alignment error rates (%) for the test set.

System	Total	Prolonging phonemes	Skipping phonemes	Repeating phrases	Late termination
Tacotron	8.5	1.6	4.8	**0.5**	1.6
Tacotron-GST-8	11.1	3.2	6.4	1.6	1.1
Tacotron-ATTR	11.6	1.1	9.0	2.7	0.5
Tacotron-context	7.4	2.1	4.2	1.1	**0.0**
Tacotron-GST-8-ATTR	11.6	2.1	8.5	1.6	1.1
Tacotron-GST-8-context	11.6	3.2	5.8	3.2	1.1
SA-Tacotron	10.1	3.2	5.8	2.1	**0.0**
SA-Tacotron-GST-8	8.5	2.1	4.8	2.1	**0.0**
SA-Tacotron-ATTR	10.1	1.6	7.4	1.6	1.1
SA-Tacotron-context	9.0	2.1	5.8	1.1	0.5
SA-Tacotron-GST-8-ATTR	**6.9**	**0.0**	**2.7**	1.0	3.2
SA-Tacotron-GST-8-context	11.1	2.7	5.3	4.2	0.5

Table 6.2: Pitch-accent error rates (%) for the test set.

System	
Tacotron	13.9
Tacotron-GST-8	14.1
Tacotron-ATTR	12.3
Tacotron-context	11.7
Tacotron-GST-8-ATTR	14.3
Tacotron-GST-8-context	14.6
SA-Tacotron	14.9
SA-Tacotron-GST-8	15.4
SA-Tacotron-ATTR	16.3
SA-Tacotron-context	**8.8**
SA-Tacotron-GST-8-ATTR	9.6
SA-Tacotron-GST-8-context	15.6

6.4 What Is Missing in Synthesized Rakugo Speech?

As mentioned in 6.2.1, Tacotron 2 can produce speech that sounds as natural as human speech in the case of well-articulated read speech. However, in the case of rakugo speech, all the TTS models including Tacotron 2 could not achieve the same scores regarding naturalness as that for AbS speech. Regarding distinguishability of characters and understandability of content, there were also significant differences between the scores for each model and AbS speech. In other words, speech synthesis currently cannot reach the professional level of rakugo performance.

The most important point of our listening test was whether the listeners were entertained by synthetic rakugo speech. However, none of the TTS models obtained scores equivalent to AbS speech. Interestingly, there were lower significant differences between the scores for some of the TTS models and AbS speech.

Unfortunately, GSTs and context embeddings did not contribute well to a higher quality of generated speech. This may be caused by the relatively small amount of speech data used for training the models. As mentioned in 6.3.2, the training set only contains a total of 3.74 hours of speech data, while the quantity of speech data used in the original GST paper is much larger [37]. We would need a much greater amount of rakugo speech or other expressive speech for properly training the GSTs

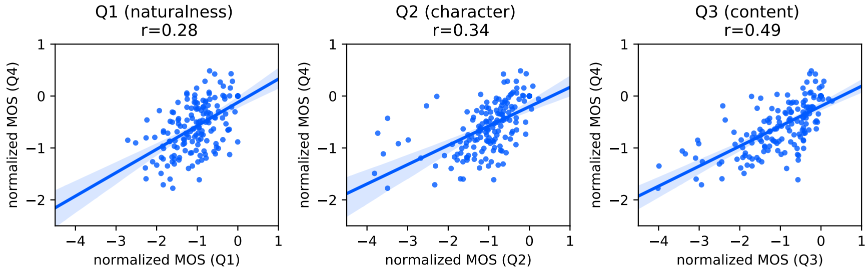

Figure 6.6: Scatter plots and correlation coefficients of normalized MOSs between each of Q1–Q3 and Q4. Blue lines and light blue areas represent simple linear regression lines and 95% confidence intervals, respectively.

and effectively modeling speaking styles.

For further analysis, we calculated the correlation among the scores for Q4, the question for evaluating the degree of entertainment, and those for the other questions. The results are shown in Figure 6.6. The correlation coefficients between the scores for Q4 and those for Q2 (distingishability of characters) and between the scores for Q4 and those for Q3 (understandability of content) were higher than the coefficient between the scores for Q4 and those for Q1 (naturalness). This suggests that we should not only focus on the naturalness of synthesized speech, but we should also aim to improve the distinguishability of characters and the understandability of the content in order to further entertain listeners. We believe that this is an important insight for the speech synthesis community since speech synthesis research has thus far mainly focused on the naturalness over other aspects.

What should be improved to further entertain listeners in particular? To investigate this, we analyzed the relationship between fundamental frequency (f_o) and speech rates of the natural speech and of each model, and the scores for each model. We extracted f_o for all the voiced frames (5-ms frame shift) in each sentence. The f_os were extracted by WORLD [69] and then corrected manually. The extracted f_os were concatenated over all the test sentences per model, and the means and standard deviations were calculated.

The relationship between the ratio of the standard deviation of f_o to its mean and the scores of each question that we evaluated are shown in Figure 6.7. We can clearly see that only Q4 (entertaining) has moderate correlation between the f_o's variations

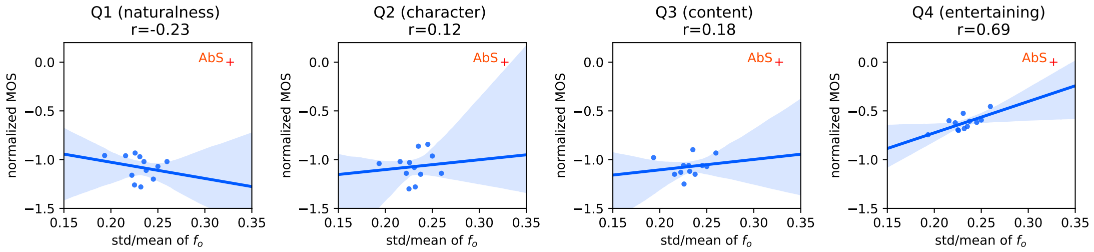

Figure 6.7: Scatter plots and correlation coefficients of ratio of standard deviation to mean of f_o and normalized MOSs. Blue lines and light blue areas represent simple linear regression lines and 95% confidence intervals, respectively. AbS (red plus marks) were not considered when calculating correlation coefficients, regression lines, and confidence intervals.

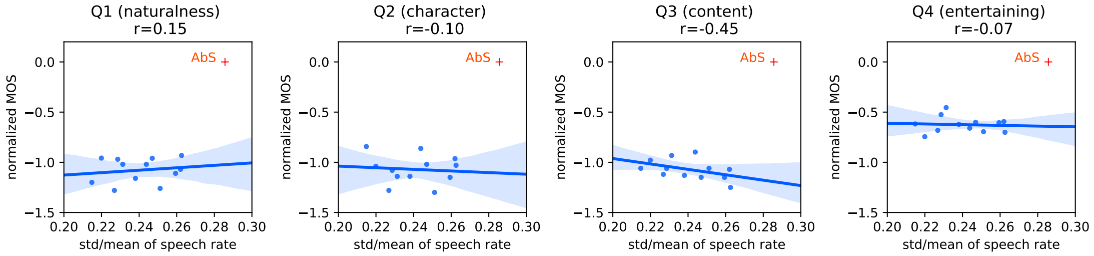

Figure 6.8: Scatter plots and correlation coefficients of the ratio of standard deviation to mean of speech rates and normalized MOSs. Blue lines and light blue areas represent simple linear regression lines and 95% confidence intervals, respectively. AbS (red plus marks) was not considered when calculating correlation coefficients, regression lines, and confidence intervals.

and its scores. This suggests that more entertaining speech should have richer f_o expression.

Speech rate is defined as the ratio of the number of mora to the duration of the speech. The means and standard deviations were calculated over all the test sentences per model. The relationship between the ratio of the standard deviation of speech rate to its mean and the scores are shown in Figure 6.8. Unlike the case of f_o, Q4 (entertaining) does not seem to correlate with the speech rate variations. Q1 (naturalness) and Q2 (distingishability of characters) have similar trends. Q3 (understandability of content) seems to have a negative correlation with the speech rate variations. However, its slope is almost flat.

6.5 Summary

In this chapter, we modeled rakugo speech using Tacotron 2 and an enhanced version of it with self-attention (SA-Tacotron) to better consider long-term dependencies, and compared their outputs. Through a listening test, we found that state-of-the-art TTS models could not reach the professional level, and there were statistically significant differences in terms of naturalness, distinguishability of characters, understandability of content, and even the degree of entertainment; nevertheless, the results of the listening test provided some interesting insights: 1) we should not focus only on naturalness of synthesized speech but also the distinguishability of characters and the understandability of the content to further entertain listeners; 2) the f_o expressivity of synthesized speech is poorer than that of human speech, and more entertaining speech should have richer f_o expression.

7 Comparison with Human Professionals

7.1 Motivation

As described in Chapter 2, professional (Edo) rakugo performers are ranked at one of three levels. In Chapter 6, we used a shin-uchi (first-rank) performer's audio recordings only as a reference for assessing our synthesizer. However, such evaluation may not be ideal since the skills of rakugo performers also vary significantly. We should compare rakugo speech synthesizers with professional rakugo performers from the three different ranks, i.e. in ascending order, zenza, futatsume, and shin-uchi. This should clarify more precisely what is missing in our rakugo synthesizer to entertain audiences.

In this chapter, we therefore propose a novel subjective evaluation methodology using natural speech uttered by performers from the three different ranks in addition to synthesized speech and show benchmarking results for our rakugo speech synthesizer. For this purpose, as described in 3.2.3, we recorded speech of a common story performed by a performer of each of the three ranks and then conducted a subjective comparison with synthesized speech of the same story through a listening test in terms of the four aspects which were also assessed in Chapter 6: 1) naturalness, 2) distinguishability of characters in the story, 3) understandability of the content, 4) degree of entertainment, and an aspect newly assessed in this chapter: 5) performer's skill level.

The content of this chapter is based on [Pub3].

7.2 Related Work

As described in 4.3.2, audiobook speech and its TTS are different from those of rakugo but related to each other. A subjective evaluation methodology for audiobook TTS is proposed in [130]. In [130], listeners are asked to answer nine questions: 1) overall impression, 2) voice pleasantness, 3) accentuation, 4) listening effort, 5) comprehension problems, 6) acceptance, 7) speech pauses, 8) intonation, and 9) emotion. In the Blizzard Challenge 2017, a competition for audiobook TTS, evaluation was conducted utilizing this methodology [131].

Although we could adopt a similar methodology to evaluate rakugo TTS, we constructed a different methodology to focus on the degree of entertainment as rakugo is authentically a form of entertainment; therefore we adopt the same questions as in Chapter 6 to measure the degree of entertainment and its possible factors, and added a question to measure the performer's skill level, which is also expected to relate to the degree of entertainment. In addition, we used natural recordings performed by performers of various levels. This design is aiming to rate rakugo TTS in the context of rank for (human) professional performers.

7.3 Natural Speech by Human Professionals Used in the Listening Test

We used natural rakugo speech in the Database II (3.2.3) to be compared with synthesized rakugo speech. As described in 3.2.3, the Database II is composed of the speech of professional performers of the three ranks. The common story is a story called "Misomame." The total durations of the recordings by the zenza, futatsume, and shin-uchi were 2.5, 2.7, and 4.2 minutes, respectively.

As described later in this chapter, speech of a story called “Misomame” was used in the listening test, although speech of shorter stories was used for the listening test in Chapter 6. The reason is that we believe a “full” story[1] is more suitable for evaluating the level of rakugo speech synthesis than a short story. On the other hand, a long story will not be suitable for a listening test considering the load and quality of evaluation. We therefore adopted Misomame, which is a full story, though relatively short in duration, on the basis of advice from Yanagiya Sanza, the shin-uchi performer above.

7.4 Synthesized Speech Used in the Listening Test

We used a variant of the Tacotron-based TTS system (SA-Tacotron-context model from Chapter 6) because this model was evaluated as the best one. This model takes textual information and context embeddings as inputs. It should be reminded that the model is based on the Database I, which is composed of speech by the shin-uchi performer. The Database II was not used for model building and was used only for comparison with synthesized speech. Minor differences from Chapter 6 were as follows. 1) The sentences in “Misomame” were excluded from the training set and validation set. As a result, we used 6,362 sentences (3.67 hours) for training, 706 sentences (0.42 hours) for validation, and 273 sentences (0.22 hours) for testing. 2) The sampling frequency was changed from 16 kHz to 24 kHz for mel-spectrogram output from the speech synthesis model and waveforms generated through a WaveNet [35] vocoder [71, 72]. Accordingly, the frame shift and fast Fourier transform size were changed to 12 ms and 2,048, respectively for better modeling.

[1] While there is no clear definition either of a full rakugo story or short story, short stories tend to appear in the makura, or prelude to the main story, of the performance of a full story, and are never independently performed on a stage. In Edo rakugo, several hundred traditional stories are performed.

7.5 Listening Test

7.5.1 Test Conditions

As mentioned in 7.3, we used speech of Misomame in the listening test[2]. The speech was synthesized sentence by sentence. Pauses between sentences were not predicted[3], and the pauses between sentences for the synthesized speech were the same as those of the real audio recording. Listeners evaluated the speech NOT sentence by sentence but as a whole story. All speech was normalized to −26 dBov over the whole story using sv56 [123].

We asked listeners to answer a five-scale MOS-based test. Listeners listened to either speech by the professional performers (zenza, futatsume, or shin-uchi) or the synthesized speech, and they evaluated them according to the five questions below.

1) How natural did the performer sound? (naturalness)

2) How accurately did you think you could distinguish each character?

3) How well did you think you could understand the content?

4) How well were you entertained?

5) How high was the rakugo skill level of the performer?

The most important question was Q4 since rakugo is a form of verbal entertainment. Q5 was intended for evaluating the "skill level" of the rakugo speech synthesis as if it were a professional performer. The others were questions about factors that we hypothesized may affect the results of Q4 and Q5. A total of 292 listeners participated in 292 evaluation rounds through crowdsourcing.

7.5.2 Results

The listening test results are shown in Figure 7.1, where SS, NS, NF, and NZ correspond to speech synthesis, shin-uchi, futatsume, and zenza, respectively. For statistical analysis, we used Brunner-Munzel tests with Bonferroni correction.

[2] The details of the test story is described in Appendix A.

[3] They should be predicted, but that is out of the scope of this thesis.

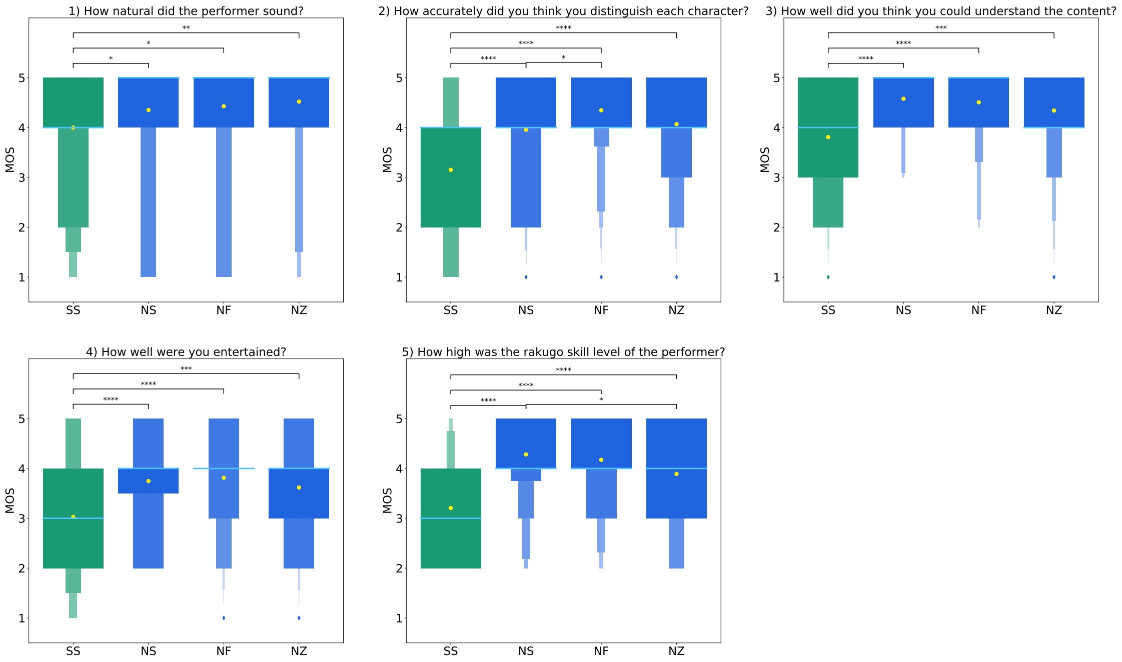

Figure 7.1: Boxen plots for each question of listening test. Light blue lines represent medians, and yellow dots represent means. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.005$, ****: $p < 0.001$.

As can be seen from the figure, the scores of the speech synthesis did not reach those of the natural speech of the professional performers, but we see that the trends in the score differences were different depending on the question.

For Q1 (naturalness), the mean score for the speech synthesis was 4.0. This means that the naturalness of the synthesized speech was high and comparable enough to that of natural speech. On the contrary, for Q2 (character), Q3 (content), and Q4 (entertaining), the mean scores for the speech synthesis ranged between 3.0 and 4.0, which were much lower than those for the professional performers. For Q3 and Q4, the p-values between the scores for the speech synthesis and those for the zenza were also smaller than the p-values between the scores for the speech synthesis and those for the futatsume or shin-uchi.

For Q5, which measures the skill level of the performer, the mean scores descended according to rank (shin-uchi > futatsume > zenza) as we expected. The synthesized speech was rated lower than the natural performances.

Table 7.1: Correlation coefficients of MOSs between questions.

	Q2	Q3	Q4	Q5 (skill level)
Q1 (naturalness)	0.287	0.303	0.317	0.339
Q2 (character)	-	0.538	0.486	0.580
Q3 (content)	-	-	0.597	0.582
Q4 (entertaining)	-	-	-	0.656

7.5.3 Correlations Among Questions

To understand the listening test results better, we calculated the correlation coefficients of the MOSs among the questions, and the results are shown in Table 7.1. We see that the Q4 (entertaining) scores had a larger correlation coefficient in the order of the scores for Q5 (skill level), Q3 (content), Q2 (character), and Q1. In other words, Q1 (naturalness) had the weakest correlation coefficient with Q4 (entertaining). The correlation coefficient between the scores for Q2 and those for Q3 was also relatively large. In summary, while the skill level (Q5), entertainment (Q4), understandability of the content (Q3), and distinguishability of the characters (Q2) were correlated with each other to a moderate degree, naturalness (Q1) appeared to be less correlated with the other metrics.

From the above results in Figure 7.1, we learned that, even though the naturalness of the synthesized rakugo speech was close to that of the human professionals, it could not sufficiently entertain the listeners because the listeners could not perfectly distinguish characters in the synthesized speech and therefore could not adequately understand the content. In other words, we should not only improve the naturalness of synthesized speech but also refine the modeling of other aspects of speech, such as the distinguishability of characters in the case of rakugo, to better entertain listeners.

7.5.4 Why Were the Degrees of Entertainment Not Assessed According to Rank?

For Q4, which measures the degree of entertainment, the mean scores did not descend according rank (futatsume > shin-uchi > zenza) against our expectation. This might caused by the selection of the listeners of the listening test. Since we did not ask the

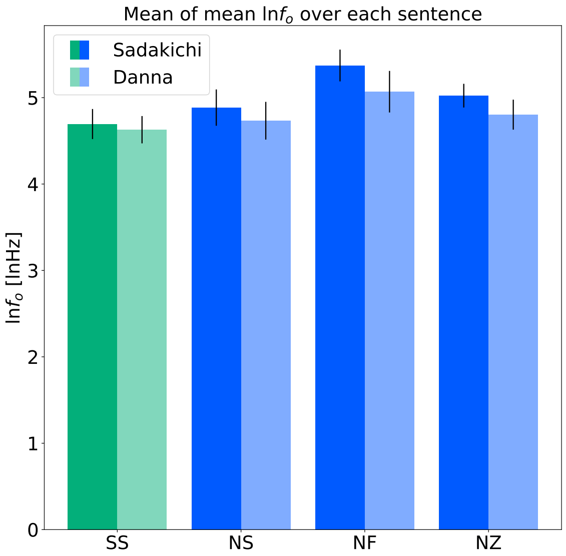

Figure 7.2: Means of means and standard deviations of logarithmic fundamental frequency (lnf_o) over each sentence. *Sadakichi* and *Danna* are two characters performed by performers.

listeners be familiar with rakugo performance, this listening test may have been the first time some listeners had ever heard rakugo considering the popularity of rakugo in Japan[4]. Such listeners would not understand the manners of rakugo performance and could not enjoy skilled performance enough.

7.6 Acoustic Analysis

We further investigated what makes it difficult for listeners to distinguish characters. We calculated the mean and standard deviation of the logarithmic fundamental frequency (lnf_o) and duration per mora, sentence by sentence, and averaged them over the story for each character. Misomame has two characters, *Sadakichi* (a boy) and *Danna* (a middle-aged male). The results for the lnf_o and duration corresponding to the two characters in the test set are shown in Figures 7.2 and 7.3.

In Figure 7.2, we can see that the degree of cross-character difference for the mean lnf_o was different according to performer (futatsume > zenza > shin-uchi >

[4] As mentioned in Chapter 2, rakugo is one of a popular traditional forms of entertainment in Japan. But its popularity is somewhat limited compared to modern forms of entertainment.

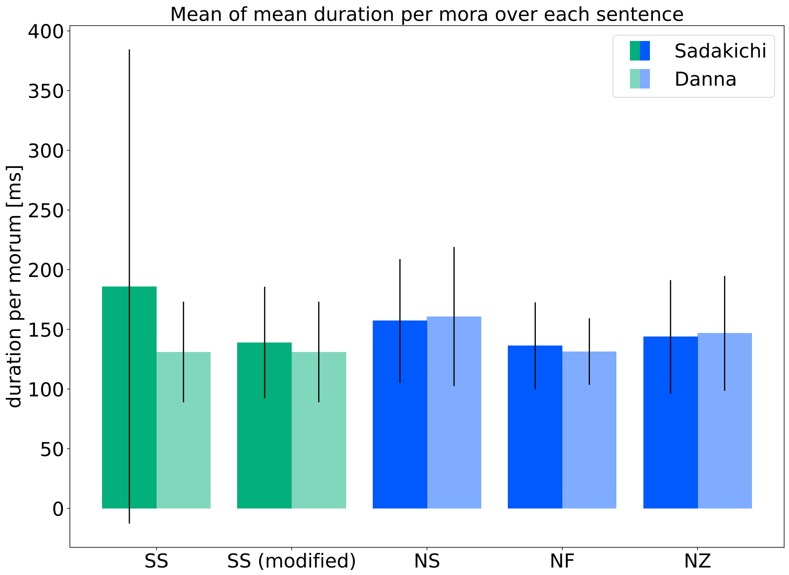

Figure 7.3: Means of means and standard deviations of duration per mora over each sentence. SS (modified) was calculated on basis of sentences, excluding two sentences for which duration was estimated as too long.

synthesized speech). We can also see that the order of the standard deviations of $\ln f_o$ was different according to performer (futatsume shin-uchi > zenza synthesized speech). Considering these facts, larger cross-character difference for the mean $\ln f_o$ and larger standard deviation of $\ln f_o$ would make it easier to distinguish characters in the story. If so, it is obvious that synthesized speech should have larger cross-character difference for the mean $\ln f_o$ and larger standard deviation of $\ln f_o$; however we should not thoughtlessly make them bigger. We need to consider that the extent to which a performer differentiates the voices of different characters depends on the performer. Yanagiya Sanza, the shin-uchi performer, does not strongly distinguish characters, according to an interview [132]. However, the difference for the synthesized speech was even smaller than that of the shin-uchi performer.

In Figure 7.3, we can see that all of the human professionals did not strongly distinguish characters using speech rates in the case of “Misomame.” This was the same for speech synthesis. We would like to note that professional performers may clearly change speech rates depending on the character [Pub4], and some synthesized speech samples used in our previous study [Pub2] distinguished characters using speech rates much more mildly than in the natural samples.

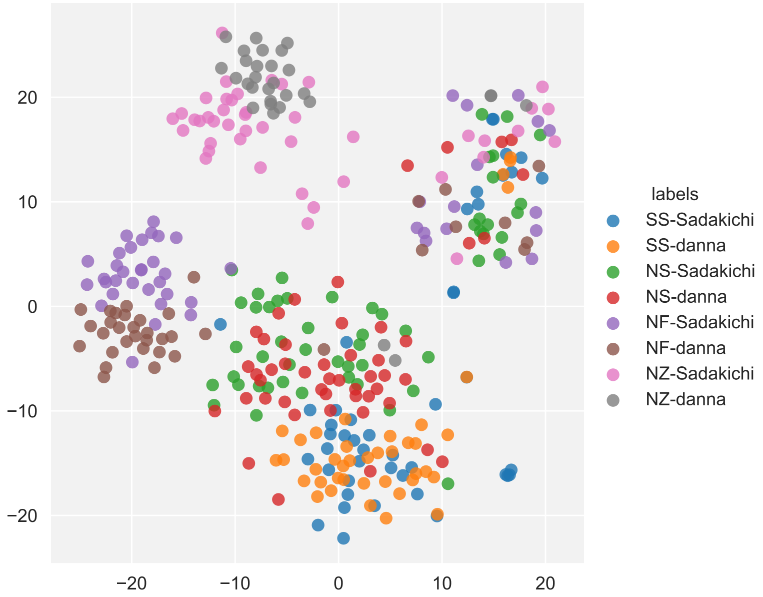

Figure 7.4: Visualization of x-vector for each sentence using t-SNE.

How about speaker individuality? Figure 7.4 is a t-SNE [133] visualization of the x-vector [134] for each sentence. We can see four clusters, namely, “zenza,” “futatsume,” “shin-uchi and speech synthesis,” and “all the systems.” The “zenza” and “futatsume” clusters were generally divided into sub-clusters by character. In comparison, the “shin-uchi” cluster did not have clear separation by character, even though the listeners could distinguish the characters according to our listening test results. We therefore consider that the shin-uchi performer used different acoustic cues to express the characters from those of the lower rank performer. The “speech synthesis” cluster did not have separation by character, either.

7.7 Summary

In this chapter, we proposed a novel methodology for evaluating rakugo speech and conducted a listening test to investigate how the level of rakugo speech synthesis compares to professional rakugo performers at various levels. From the listening test results, we found that the level of speech synthesis did not reach that of human professionals; nevertheless, the results suggested that we should make the f_o expression of speech synthesis richer to better entertain audiences. This suggestion strengthens the conclusions in Chapter 6.

8
Conclusion

8.1 Replies to the Issues

Issue 1: There is no usable rakugo speech databases for speech synthesis.

In Chapter 3, we built the first rakugo speech database usable for TTS. We not only transcribed the speech but also appended context labels manually for each sentence.

Issue 2: Rakugo speech is far more diverse and casually-pronounced than speech ordinarily used for building speech synthesis.

We successfully modeled rakugo speech using end-to-end/sequence-to-sequence TTS. In Chapter 5, we attempted to model rakugo speech with the SSNT-based model, which has no soft attention networks, to aim to deal with the diversity and casualness of rakugo speech. We also used GSTs or manually labeled context features to enrich speaking styles of rakugo speech. The models synthesized somewhat natural, character-distinguishable, and content-understandable speech. However, the MOSs for the speech were just around 3 through a listening test. In Chapter 6, for further improvement, we replaced the SSNT-based model above to Tacotron 2, which is a state-of-the-art TTS model, or an enhanced version of it with self-attention (SA-Tacotron) to better consider long-term dependencies. We confirmed that the models synthesized speech of better quality than the SSNT-based models.

Issue 3: Characters should be easily distinguishable and contents should be easily understandable.

In Chapter 6, through a listening test, we found that state-of-the-art TTS models could not yet reach the professional level, and there were statistically significant differences in terms of naturalness, distinguishability of characters, understandability of content, and even the degree of entertainment; nevertheless, the results of the listening test provided some interesting insights: 1) we should not focus only on naturalness of synthesized speech but also the distinguishability of characters and the understandability of the content to further entertain listeners; 2) the f_o expressivity of synthesized speech is poorer than that of human speech, and more entertaining speech should have richer f_o expression.

Issue 4: Synthesized speech should entertain listeners.

In Chapter 7, we proposed a novel methodology for evaluating rakugo speech and conducted a listening test to investigate how the level of rakugo speech synthesis compares to professional rakugo performers at various levels. From the listening test results, we found that the level of speech synthesis did not reach that of human professionals; nevertheless, the results suggested that we should make the f_o expression of speech synthesis richer to better entertain audiences. This suggestion strengthens the reply to Issue 3.

8.2 Remaining Issues

Remaining issue 1: The expression of speech should be further improved.

We found that not only naturalness but also distinguishability of characters and understandability of content should be considered to better entertain audiences.

Considering the suggestions we obtained in Chapters 6 and 7, we should at least enrich f_o expressivity of synthesized speech; however, it is obvious that we should not focus only on f_o expressivity but also other possible factors that affect the degree of entertainment. We need to investigate what the hidden key factors are to improve the quality of synthesized speech.

When trying to improve distinguishability of characters, we should consider that rakugo is played by a single performer and he or she may not strongly distinguish each character according to his or her style of performance. In addition, the frequency of the properties of the characters (gender, age, social rank, etc.) in common rakugo stories is very unbalanced. For example, young townsmen appear in rakugo stories very frequently, and women servants to samurai warriors rarely appear. We should consider such the difficulties when designing a model.

We could also improve the expression in other ways such as considering over-sentence dependency. Speech synthesis systems basically synthesize speech sentence by sentence. The models we built in this thesis also did so; rakugo performers, however, definitely produce speech of a sentence considering over-sentence dependency. To produce speech of a sentence, they may consider before and after some sentences, or even consider ochi (the punch line of a story).

Remaining issue 2: Duration of pauses between sentences should be estimated.

In this thesis, we did not estimate the duration of pauses between sentences and just used the pause durations found in natural recordings. Needless to say, this has to be estimated to realize fully-synthesized rakugo speech synthesis.

Remaining issue 3: We miss authentic elements which human rakugo performers have or use.

Rakugo synthesizers we built in this thesis form a first step towards realizing an authentic rakugo synthesizer. Such an authentic rakugo synthesizer should

- perform considering reactions and feedback from an audience;
- generate new rakugo stories;
- not only perform verbally but also perform visually.

Publications

[Pub1] Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, Shinji Takaki, and Junichi Yamagishi, “Rakugo speech synthesis using segment-to-segment neural transduction and style tokens — toward speech synthesis for entertaining audiences,” in *Proc. The 10th ISCA Speech Synthesis Workshop (SSW10)*, Wien, Austria, Sep 20–22 2019, pp. 111–116, doi: 10.21437/SSW.2019-20.

[Pub2] ——, “Modeling of rakugo speech and its limitations: Toward speech synthesis that entertains audiences,” *IEEE Access*, vol. 8, pp. 138 149–138 161, Jul 27 2020, doi: 10.1109/ACCESS.2020.3011975.

[Pub3] Shuhei Kato, Yusuke Yasuda, Xin Wang, Erica Cooper, and Junichi Yamagishi, “How similar or different is rakugo speech synthesizer to professional performers?” in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Jun 6–11 2021.

[Pub4] Shuhei Kato, Shinji Takaki, Junichi Yamagishi, and Xin Wang, “Investigation of rakugo speech synthesis and analysis of context using WaveNet: Toward speech synthesis entertaining people (in Japanese),” in *Proc. 2018 ASJ Autumn Meeting*, Oita, Oita, Japan, 2018, pp. 1139–1142.

Bibliography

[1] Suzumoto Engeijo, 7-12, Ueno 2-chome, Taito, Tokyo, Japan, 1857–present. [Online]. Available: http://www.rakugo.or.jp

[2] Suehirotei, 6-12, Shinjuku 3-chome, Shinjuku, Tokyo, Japan, 1897–present. [Online]. Available: http://www.suehirotei.com

[3] Asakusa Engei Hall, 43-12, Asakusa 1-chome, Taito, Tokyo, Japan, 1964–present. [Online]. Available: http://www.asakusaengei.com

[4] Ikebukuro Engeijo, 23-1, Nishi-ikebukuro 1-chome, Toshima, Tokyo, Japan, 1951–present. [Online]. Available: http://www.ike-en.com

[5] Shumputei Shotaro, Aug 23 1981–present. [Online]. Available: http://shoutarou.com

[6] This photo is transformed from "DP3M2471" by akira kawamura licensed under CC BY 2.0.

[7] Avaya Inc., "Automatic Solutions for Contact Centers," 2021. [Online]. Available: https://www.avaya.com/en/products/contact-center/voice/

[8] Cisco Systems, Inc., "Cisco Unified IP Interactive Voice Response (IVR)," 2021. [Online]. Available: https://www.cisco.com/c/en/us/products/contact-center/unified-ip-interactive-voice-response-ivr/

[9] Concentrix Corporation, "Conversational AI," 2021. [Online]. Available: https://www.concentrix.com/solutions/digital-self-service/conversational-ai/

[10] 8x8, Inc., "IVR Software," 2021. [Online]. Available: https://www.8x8.com/products/contact-center/intelligent-ivr

[11] Nuance Communications, Inc., "Conversational IVR Technology For Customer Self Service," 2021. [Online]. Available: https://www.nuance.com/omni-channel-customer-engagement/voice-and-ivr/conversational-ivr.html

[12] Cerence, Inc., "Products," 2021. [Online]. Available: https://www.cerence.com/cerence-products/overview

[13] ReadSpeaker Holding B.V., "Automotive and Text to Speech," 2021. [Online]. Available: https://www.readspeaker.com/blog/industries/automotive/

[14] Amazon.com, Inc., "Amazon Alexa Official Site: What is Alexa?" 2021. [Online]. Available: https://developer.amazon.com/en-US/alexa

[15] Apple Inc., "Siri," 2021. [Online]. Available: https://www.apple.com/siri/

[16] Google LLC, "Assistant, your own personal Google," 2021. [Online]. Available: https://assistant.google.com

[17] Microsoft Corporation, "Cortana - Your personal productivity assistant," 2021. [Online]. Available: https://www.microsoft.com/en-us/cortana

[18] Google LLC, "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone," 2018. [Online]. Available: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

[19] LINE Corporation, "Line AiCall," 2021. [Online]. Available: https://clova.line.me/line-aicall/

[20] Amazon.com, Inc., "Books on Alexa," 2021. [Online]. Available: https://www.amazon.com/gp/browse.html?node=17373083011

[21] Apple Inc., "Accessibility - Vision," 2021. [Online]. Available: https://www.apple.com/accessibility/vision/

[22] Freedom Scientific Inc., "JAWS Headquarters," 2021. [Online]. Available: https://support.freedomscientific.com/JAWSHQ/JAWSHeadquarters01

[23] Google LLC, "Turn on TalkBack - Android Accessibility Help," 2021. [Online]. Available: https://support.google.com/accessibility/answer/7031755

[24] ——, “Use the built-in screen reader - Google Accessibility Help,” 2021. [Online]. Available: https://support.google.com/accessibility/android/answer/6007100

[25] Microsoft Corporation, “Complete guide to Narrator,” 2021. [Online]. Available: https://support.microsoft.com/en-us/help/22798/windows-10-complete-guide-to-narrator

[26] NV Access Limited, “About NVDA,” 2021. [Online]. Available: https://www.nvaccess.org/about-nvda/

[27] HOYA Corporation, “Show-Kun Kona (in Japanese),” 2019. [Online]. Available: https://readspeaker.jp/show/

[28] NHK, “Yomiko No Heya (in Japanese),” 2021. [Online]. Available: https://www.nhk.or.jp/voice/yomiko/

[29] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr 15–20 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.

[30] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, “Neural speech synthesis with transformer network,” in *Proc. AAAI Conf. Artif. Intell. (AAAI-19)*, Honolulu, HI, USA, Jan 27 – Feb 1 2019, doi: 10.1609/aaai.v33i01.33016706.

[31] Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, and Aaron C. Courville, “Char2Wav: End-to-end speech synthesis,” in *Proc. Intl. Conf. Learn. Representations (ICLR)*, Palais des Congrès Neptune, Toulon, France, Apr 24–27 2017.

[32] Sercan O. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi, “Deep Voice: Real-time neural text-to-speech,” in *Proc. Intl. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug 6–11 2017.

[33] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Stockholm, Sweden, Aug 20–24 2017, pp. 4006–4010, doi: 10.21437/Interspeech.2017-1452.

[34] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani, "VoiceLoop: Voice fitting and synthesis via a phonological loop," in *Proc. Intl. Conf. Learn. Representations (ICLR)*, Vancouver, BC, Canada, Apr 30 – May 3 2018.

[35] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," arXiv:1609.03499 [cs.SD], Sep 12 2016.

[36] Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A. Saurous, "Uncovering latent style factors for expressive speech synthesis," in *Proc. Conf. Neural Inform. Process. Syst. (NIPS) Mach. Learn. for Audio Signal Process. Workshop*, Long Beach, CA, USA, Dec 4–9 2017, pp. 4006–4010.

[37] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Intl. Conf. Mach. Learn. (ICML)*, Stockholm, Stockholm, Sweden, Jul 10–15 2018, pp. 5180–5189.

[38] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," arXiv:1803.09047 [cs.CL], Mar 24 2018.

[39] Trevor Wood, "Varying speaking styles with neural text-to-speech," Alexa Blogs, Nov 19 2018. [Online]. Available: https://developer.amazon.com/zh/blogs/alexa/post/7ab9665a-0536-4be2-aaad-18281ec59af8/varying-speaking-styles-with-neural-text-to-speech

[40] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. Intl. Conf. Learn. Representations (ICLR)*, New Orleans, LA, USA, May 6–9 2019.

[41] Sean Vasquez and Mike Lewis, "MelNet: A generative model for audio in the frequency domain," arXiv:1906.01083 [eess.AS], Jun 4 2019.

[42] Eric Battenburg, Soroosh Mariooryad, Daisy Stanton, RJ Skerry-Ryan, Matt Shannon, David Kao, and Tom Bagby, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," arXiv:1906.03402 [cs.CL], Jun 8 2019.

[43] Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, David Kao, and Tom Bagby, "Semi-supervised generative modeling for controllable speech synthesis," arXiv:1910.01709 [cs.CL], Oct 3 2019.

[44] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 4–8 2020, pp. 6264–6268, doi: 10.1109/ICASSP40776.2020.9053520.

[45] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 4–8 2020, pp. 6699–6703, doi: 10.1109/ICASSP40776.2020.9053436.

[46] MSS, "Hanashika Miku No Tokusen Rakugo Manju Kowai Desu (in Japanese)," Jan 20 2009. [Online]. Available: https://www.nicovideo.jp/watch/sm5899050

[47] Metsuki-warui-P, "[VOCALOID Rakugo] Kamban No Pin [Metsuki Warui Miku] (in Japanese)," Mar 25 2011. [Online]. Available: https://www.nicovideo.jp/watch/sm13959846

[48] zky, "[Hatsune Miku] VOCALOID Rakugo "Nozarashi" (in Japanese)," Feb 24 2012. [Online]. Available: http://www.nicovideo.jp/watch/sm17066984

[49] Yamamoto Susumu, *Tanoshii Rakugo: Edo Irai Yonhyaku-nen, Soshite Mirai E (in Japanese).* Shinjuku, Tokyo, Japan: Soshisha, Dec 18 2013.

[50] Atoss Broadcasting Limited, "Yose Channel," Oct 1 2012–present. [Online]. Available: http://yosechannel.com

[51] Chiba Television Broadcasting Corporation, "Asakusa Ochanoma Yose," Apr 5 2004–present. [Online]. Available: https://www.chiba-tv.com/program/detail/1012

[52] NHK, "Kamigata Raukgo No Kai," Apr 24 2011–present. [Online]. Available: https://www4.nhk.or.jp/P2851/

[53] ——, "Shin-uchi Kyoen," Nov 26 1978–present. [Online]. Available: https://www4.nhk.or.jp/P632/

[54] Nippon Cultural Broadcating Incorporated, "Shinosuke Radio Rakugo De Date," Apr 7 2007–present. [Online]. Available: http://www.joqr.co.jp/blog/rakugo/

[55] Nikkei Radio Broadcasting Corporation, "Yose Apuri — Warai Suginami Yose Kara," Oct 22 2017–present. [Online]. Available: http://www.radionikkei.jp/yose/

[56] TBS Radio Incorporated, "Radio Yose," Oct 1974–present. [Online]. Available: https://www.tbsradio.jp/yose/

[57] Masaaki Nomura, *Rakugo No Gengogaku (in Japanese).* Bunkyo, Tokyo, Japan: Heibonsha, May 18 1994.

[58] Shinpei Takahashi and Hikaru Inooka, "Generation of animated motion fitted with talk in comic storytelling (in Japanese)," *J. Jpn. Soc. Kansei Eng.*, vol. 5, no. 1, pp. 1–6, 2004, doi: 10.5057/jjske2001.5.1.

[59] Hiroaki Kawashima, Takeshi Nishikawa, and Takashi Matsuyama, "Analysis of visual timing structure in rakugo turn-taking (in Japanese)," *IPSJ J.*, vol. 48, no. 12, pp. 3715–3728, 2007.

[60] Yanagiya Sanza, Jul 4 1974–present. [Online]. Available: http://www.yanagiya-sanza.com(inJapanese)

[61] Wolfgang von Kempelen, *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine (in Germany).* Wien, Austria: J. B. Degen, 1791.

[62] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eur. Conf. Speech Commun. and Tech. (EUROSPEECH)*, Budapest, Central Hungary, Hungary, Sep 5–9 1999, pp. 2347–2350.

[63] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 26–31 2013, pp. 7962–7966, doi: 10.1109/ICASSP.2013.6639215.

[64] Amazon.com, Inc., "Amazon Polly," 2021. [Online]. Available: https://aws.amazon.com/polly/

[65] Google LLC, "Text-to-Speech: Lifelike Speech Synthesis," 2021. [Online]. Available: https://cloud.google.com/text-to-speech/

[66] Satoshi Imai, Kazuo Sumita, and Chieko Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis (in Japanese)," *Electron. and Commun. in Jpn. (Part I: Commun.)*, vol. 66, no. 2, pp. 10–18, 1983, doi: 10.1002/ecja.4400660203.

[67] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alainde Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999, doi: 10.1016/S0167-6393(98)00085-5.

[68] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nishimura, Toshio Irino, and Hideki Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Las Vegas, NV, USA, Mar 31 – Apr 4 2008, pp. 3933–3936, doi: 10.1109/ICASSP.2008.4518514.

[69] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inform. and Syst.*, vol. E99-D, no. 7, pp. 1877–1884, 2016, doi: 10.1587/transinf.2015EDP7457.

[70] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 15 1997, doi: 10.1162/neco.1997.9.8.1735.

[71] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, and Junichi Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr 15–20 2018, pp. 4804–4808, doi: 10.1109/ICASSP.2018.8461452.

[72] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, Stockholm, Stockholm, Sweden, Aug 20–24 2017, pp. 1118–1122, doi: 10.21437/Interspeech.2017-314.

[73] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. Intl. Conf. Mach. Learn. (ICML)*, Stockholm, Stockholm, Sweden, Jul 10–15 2018.

[74] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Brighton, England, UK, May 12–17 2019, pp. 3617–3621, doi: 10.1109/ICASSP.2019.8683143.

[75] Jean-Marc Valin and Jan Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Brighton, England, UK, May 12–17 2019, pp. 5891–5895, doi: 10.1109/ICASSP.2019.8682804.

[76] Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 402–415, Nov 28 2018, doi: 10.1109/TASLP.2019.2956145.

[77] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao, “Non-autoregressive neural text-to-speech,” in *Proc. Intl. Conf. Mach. Learn. (ICML)*, Jul 13–18 2020, pp. 10 192–10 204.

[78] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “FastSpeech: Fast, robust and controllable text to speech,” in *Proc. Conf. Neural Inform. Process. Syst. (NurIPS)*, Vancouver, BC, Canada, Dec 8–14 2019, pp. 3171–3180.

[79] Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao, “Flow-TTS: A non-autoregressive network for text to speech based on flow,” in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, May 4–8 2020, pp. 7209–7213, doi: 10.1109/ICASSP40776.2020.9054484.

[80] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “FastSpeech 2: Fast and high-quality end-to-end text to speech,” arXiv:2006.04558 [eess.AS], Jun 8 2020.

[81] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron Weiss, and Yonghui Wu, “Parallel Tacotron: Non-autoregressive and controllable TTS,” arXiv:2010.11439 [cs.SD], Oct 22 2020.

[82] Tao Wang, Xuefei Liu, Jianhua Tao, Jiangyan Yi, Ruibo Fu, and Zhengqi Wen, “Non-autoregressive end-to-end TTS with coarse-to-fine decoding,” in *Proc. INTERSPEECH*, Oct 25–29 2020, pp. 3984–3988, doi: 10.21437/Interspeech.2020-1662.

[83] Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa, “Investigating accuracy of pitch-accent annotations in neural network-based speech synthesis and denoising effects,” in *Proc. INTERSPEECH*, Hyderabad, Telangana, India, Sep 2–6 2018, doi: 10.21437/Interspeech.2018-1227.

[84] Hatsune Miku, Aug 31 2007–present. [Online]. Available: https://piapro.net/intl/en.html

[85] MIKU EXPO 2014 in Indonesia, Jakarta, Indonesia, May 28–29 2014. [Online]. Available: https://mikuexpo.com/indonesia_top/

[86] MIKU EXPO 2014 in Los Angeles, Los Angeles, LA, USA, Oct 11–12 2014. [Online]. Available: https://mikuexpo.com/la/

[87] MIKU EXPO 2014 in New York, New York, NY, USA, Oct 9–19 2014. [Online]. Available: https://mikuexpo.com/ny/

[88] MIKU EXPO 2015 in Shanghai, Shanghai, China, Jun 27–28 2015. [Online]. Available: https://mikuexpo.com/shanghai/

[89] MIKU EXPO 2016 Japan Tour, Fukuoka, Fukuoka, Japan; Osaka, Osaka, Japan; Nagoya, Aichi, Japan; Sapporo, Hokkaido, Japan; and Koto, Tokyo, Japan, Mar 23–24, 29, 31, Apr 5, 9–10 2016, respectively. [Online]. Available: https://mikuexpo.com/jp2016/

[90] MIKU EXPO 2016 North America, Seattle, WA, USA; San Francisco, CA, USA; Los Angeles, CA, USA; Dallas, TX, USA; Houston, TX, USA; Toronto, ON, Canada; Chicago, IL, USA; New York, NY, USA; Monterrey, Nuevo León, Mexico; and Mexico City, Mexico, Mexico, Apr 23, 30, May 6, 14, 17, 20, 25, 28, Jun 1, and 4–5 2016, respectively. [Online]. Available: https://mikuexpo.com/na2016/

[91] MIKU EXPO 2016 in Taiwan, New Taipei City, Northern Taiwan, Taiwan, Jun 25–26 2016. [Online]. Available: https://mikuexpo.com/tw2016/

[92] MIKU EXPO 2016 China Tour, Shanghai, China; and Beijing, China, Dec 3–4, and 10–11 2016, respectively. [Online]. Available: https://mikuexpo.com/cn2016/

[93] MIKU EXPO 2017 in Malaysia, Kuala Lumpur, Malaysia, Dec 16 2017. [Online]. Available: https://mikuexpo.com/may2017/

[94] MIKU EXPO 2018 USA & Mexico, Los Angeles, CA, USA; San Jose, CA, USA; Dallas, TX, USA; Austin, TX, USA; Washington, D.C., USA; New York, NY, USA; and Mexico City, Mexico, Mexico, Jun 29, Jul 1, 6, 8, 12, 14, and 19 2018, respectively. [Online]. Available: https://mikuexpo.com/usamx2018/

[95] MIKU EXPO 2018 EUROPE, Paris, Île-de-France, France; Cologne, Cologne, North Rhine-Westphalia, Germany; and London, England, UK, Dec 1, 4, and 8 2018, respectively. [Online]. Available: https://mikuexpo.com/europe2018/

[96] MIKU EXPO 2019 Taiwan & Hong Kong, New Taipei City, Northern Taiwan, Taiwan; and Hong Kong, China, May 11 and Jul 27 2019, respectively. [Online]. Available: https://mikuexpo.com/twhk2019/

[97] MIKU EXPO 2020 EUROPE, London, England, UK; Paris, Île-de-France, France; Berlin, Germany; Amsterdam, North Holland, Netherlands; and Barcelona, Catalonia, Spain, Jan 11, 16, 20, 24, and 28 2020, respectively. [Online]. Available: https://mikuexpo.com/europe2020/

[98] NHK, "NHKスペシャル「AIでよみがえる美空ひばり」," Sep 29 2019. [Online]. Available: https://www.nhk.or.jp/docudocu/program/46/2586133/index.html

[99] Masato Tanii, "「AI美空ひばり」の舞台裏「冗談でやっていいことではない」——故人をよみがえらせたヤマハの技術者の思い," ITmedia NEWS, Oct 3 2019. [Online]. Available: https://www.itmedia.co.jp/news/articles/1910/02/news076.html

[100] Makiko Yabe, "「号泣です」「涙腺崩壊」……"AI美空ひばり"が歌う新曲『あれから』に、なぜ感情を揺さぶられるのか," Bunshun Online, Oct 12 2019. [Online]. Available: https://bunshun.jp/articles/-/14618

[101] NHK, "美空ひばりが紅白で、復活！," Dec 31 2019. [Online]. Available: https://www.nhk.or.jp/kouhaku/topics/topics_191114-1.html

[102] SynSIG, "Blizzard Challenge 2012," 2012. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2012

[103] ——, "Blizzard Challenge 2013," 2013. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2013

[104] ——, "Blizzard Challenge 2015," 2015. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2015

[105] ——, "Blizzard Challenge 2016," 2016. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2016

[106] ——, "Blizzard Challenge 2017," 2017. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2017

[107] ——, “Blizzard Challenge 2018,” 2018. [Online]. Available: https://www.synsig.org/index.php/Blizzard_Challenge_2018

[108] Yusuke Yasuda, Xin Wang, and Junichi Yamagishi, “Initial investigation of an encoder-decoder end-to-end TTS framework using marginalization of monotonic hard latent alignments,” in *Proc. The 10th ISCA Speech Synthesis Workshop (SSW10)*, Vienna, Austria, Sep 20–22 2019, pp. 211–216, doi: 10.21437/SSW.2019-38.

[109] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ö. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, “Deep Voice 3: Scaling text-to-speech with convolutional sequence learning,” in *Proc. Intl. Conf. Learn. Representations (ICLR)*, Vancouver, BC, Canada, Apr 30 – May 3 2018.

[110] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. Intl. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 7–9 2015.

[111] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai, “Forward attention in sequence-to-sequence acoustic modeling for speech synthesis,” in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr 15–20 2018, pp. 4789–4793, doi: 10.1109/ICASSP.2018.8462020.

[112] Alex Graves, “Generating sequences with recurrent neural networks,” arXiv:1308.0850 [cs.NE], Aug 4 2013.

[113] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, “Attention-based models for speech recognition,” in *Proc. Conf. Neural Inform. Process. Syst. (NIPS)*, Montréal, QC, Canada, Dec 7–12 2015, pp. 577–585.

[114] Jianpeng Cheng, Li Dong, and Mirella Lapata, “Long short-term memory-networks for machine reading,” in *Proc. Conf. Empirical Methods in Natural Lang. Process. (EMNLP)*, Austin, TX, USA, Nov 1–5 2016, pp. 551–561, doi: 10.18653/v1/D16-1053.

[115] Lei Yu, Jan Buys, and Phil Blunsom, "Online segment to segment neural transduction," in *Proc. Conf. Empirical Methods in Natural Lang. Process. (EMNLP)*, Austin, TX, USA, Nov 1–5 2016, pp. 1307–1316, doi: 10.18653/v1/D16-1138.

[116] Vincent Dumoulin and Francesco Visin, "A guide to convolution arithmetic for deep learning," arXiv:1603.07285 [stat.ML], Mar 23 2016.

[117] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Intl. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 7–9 2015.

[118] Mike Schuster and Kuldip K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov 1997, doi: 10.1109/78.650093.

[119] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun 14 2014.

[120] David Krueger, Tegan Maharaj, János Kramár, Mohammad Pezeshki, Nicolas Ballas, Nan Rosemary Ke, Anirudh Goyal, Yoshua Bengio, Aaron Courville, and Chris Pal, "Zoneout: Regularizing RNNs by randomly preserving hidden activations," in *Proc. Intl. Conf. Learn. Representations (ICLR)*, Palais des Congrès Neptune, Toulon, France, Apr 24–27 2017.

[121] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods in Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct 25–29 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[122] Hisashi Kawai, Tomoki Toda, Junichi Yamagishi, Toshio Hirai, Jinfu Ni, Nobuyuki Nishizawa, Minoru Tsuzaki, and Keiichi Tokuda, "XIMERA: A concatenative speech synthesis system with large scale corpora (in Japanese)," *IEICE Trans. Inform. and Syst. (Japanese Ed.)*, vol. 89, no. 12, pp. 2688–2698, 2006.

[123] International Telecommunication Union, Recommendation G.191: Software Tools and Audio Coding Standardization, Nov 11 1993.

[124] Edgar Brunner and Ullrich Munzel, "The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation," *Biometrical J.*, vol. 42, no. 1, pp. 17–25, Jan 2000, doi: 10.1002/(SICI)1521-4036(200001)42:1%3C17::AID-BIMJ17%3E3.0.CO;2-U.

[125] Yamagishi Laboratory of the National Institute of Informatics, "self-attention-tacotron," Nov 5 2018–present. [Online]. Available: https://github.com/nii-yamagishilab/self-attention-tacotron

[126] ——, "tacotron2," Apr 29 2018–present. [Online]. Available: https://github.com/nii-yamagishilab/tacotron2

[127] Yusuke Yasuda, Xin Wang, and Junichi Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Brighton, England, UK, May 12–17 2019, pp. 6905–6909, doi: 10.1109/ICASSP.2019.8682353.

[128] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. INTERSPEECH*, Graz, Styria, Austria, Sep 15–19 2019, pp. 1526–1530, doi: 10.21437/Interspeech.2019-2441.

[129] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," arXiv:1706.03762 [cs.CL], Jun 12 2017.

[130] Florian Hinterleitner, Georgina Neitzel, Sebastian Möller, and Christoph Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Proc. Blizzard Challenge*, 2011.

[131] Simon King, Lovisa Wihlborg, and Wei Guo, "The Blizzard Challenge 2017," in *Proc. Blizzard Challenge*, 2017.

[132] Tokyo Bar Association, "Interview: Yanagiya Sanza, a professional rakugo performer (in Japanese)," *LIBRA*, vol. 11, no. 11, pp. 22–25, 2011.

[133] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov 2008.

[134] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr 15–20 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

A

Details of the Test Set

A.1 List of the Test Stories Used in Chapters 5 and 6

We used 12 and 13 stories in Chapters 5 and 6, respectively. The details of the test stories are described below.

A.1.1 #1 Story About Foolish Brothers

From: Makura of 道具屋 (*Doguya*)

Used in: Both Chapters 5 and 6

Performer (P)	アー昔からこの、オーぼんやりしたところを扱った小噺なんというのがありまして、エー兄弟が、アふたーり揃って馬鹿だったってお話が残っておりまして。
Younger brother (YB)	あんちゃん。
——	あんちゃん。
Elder brother (EB)	なんだ弟。
YB	あんちゃんのめえだけど、あの一年てえのは、エー十三か月だよね。
EB	お前は馬鹿だね。
——	そういうこと言ってるから世間でなんだかんだイ言われんだよーく覚えときな一年というのは、十四か月だで。
YB	そうなの？
——	だってあたいゆんべ、エー、よく勘定したんだよ。
——	一月でしょ？
——	二月でしょ？
——	三月四月五月六月七月八月九月十月、十一月十二月、お正月ってほら、十三か月じゃねえか。
EB	馬鹿。
——	お盆が抜けてるじゃねえか。
P	なんてんで、エーこれどっちが抜けてんだかよく分かりませんが。

A.1.2 #2 Story About Foolish Family

From: Makura of 道具屋 (*Doguya*)

Used in: Both Chapters 5 and 6

Note: Originally continuation of Story #1

Performer (P)	デでこれが一番すごいのかなと思ったら、アーもう一杯上手が、アーあったんすな。
Younger brother (YB)	はあーあんちゃん。
——	あんちゃん。
Elder brother (EB)	なんだ弟。
YB	あんちゃんの前だけど、あの来年の三月の女の子のお節句と、五月の男の子のお節句は、あの、どっちが、先に来るの？
EB	そりゃお前、三月のお節句が先に来ることもあれば、五月のお節句が先に来ることも、ああるよ。
——	なあおとっつぁんそうだよな。
P	つったら親父が。
Father	馬鹿。
——	来年のことが今から分かるか。
P	て。
——	おっかさんが横でこれ聞いてて。
Mother	まあやっぱりウチの人は頭がいいわ。
P	て、親子全員馬鹿だったって話が残ってますが。

A.1.3 #3 Story About Foolish Brothers

From: Makura of 金明竹 (*Kimmeichiku*)

Used in: Both Chapters 5 and 6

Note: Almost the same as Story #1

Performer (P)	エーなんですかね、エーゴ、兄弟が二人揃って馬鹿だったなんてつまんねえ話が残っておりまして。
Younger brother (YB)	あーんちゃん、ン、フ、フ、あんちゃん？
Elder brother (EB)	なんだ弟。
YB	あんちゃんのめえだけど、一年っていうのは、十三か月だよな？
EB	お前はそういうこと言っているから馬鹿にされるんだよーく覚えときな一年てえのは、十四か月だよ。
YB	そうかなあ。
——	だってあたいゆんべよーく勘定したんだよ？
——	一月でしょ？
——	二月、三月四月五月六月七月八月九月十月、十一月十二月、お正月ってほら、十三か月でしょ？
EB	馬鹿。
——	お盆が抜けてるじゃねえか。
P	エーこれどっちが抜けてるんだかよく分かりませんが。

A.1.4 #4 Story About Foolish Family

From: Makura of 金明竹 (*Kimmeichiku*)

Used in: Both Chapters 5 and 6

Note: Originally continuation of Story #3, almost the same as Story #2

Performer (P)	これが一番すごいのかと思ったら、もう一杯上手ってえのがありました。
Younger brother (YB)	あんちゃん、フッ。
——	あんちゃん。
Elder brother (EB)	なんだ弟。
YB	ア、あんちゃんの前だけど、来年の三月の女の子のお節句と五月の男の子のお節句って、どっちが先に来るの？
EB	そらあお前、そらあ三月のお節句が先に来ることもあれば、五月のお節句が先に来ることもあるよ。
——	そうだよなおとっつぁん。
P	つったら親父が。
Father	馬鹿。
——	来年のことが今から分かるか。
P	て。
——	おっかさんが横でこれ聞いてて。
Mother	まあやっぱりウチの人は頭がいいわ。
P	て親子全員馬鹿だったって話が残っておりますが。

A.1.5 #5 Story About Size of Mouse

From: Makura of 味噌豆 (*Misomame*)

Used in: Both Chapters 5 and 6

Young man #1 (YM1)	おーっと、ネズミ捕まえたネズミねえ？
——	このネズミ、ねえ？
——	これ大きいよこれ。
Young man #2 (YM2)	うーん尻尾しか見えねえけど、小さいんじゃないの？
YM1	いや大きいよ。
YM2	小さいよ。
YM1	大きい。
YM2	小さい。
Performer	たらネズミが、チュウッつったってんですな、ええ。

A.1.6 #6 Story About Drunk Crab

From: Makura of 味噌豆 (*Misomame*)

Used in: Both Chapters 5 and 6

Young man	オー見なよ見なよ見なよエエ？
——	この蟹。
——	この蟹おかしいよ？
——	蟹ってのは横に這うだろ？
——	縦に歩ってんだいどうしたんだろうね。
Performer	たら蟹が顔上げて。
Crab	アすいません、エちょっと酔ってるもんですから。

A.1.7 #7 Story About Hard of Hearing Elderly Couple

From: Makura of 味噌豆 (*Misomame*)

Used in: Both Chapters 5 and 6

Performer	耳の遠いおじいさんとおばあさんの会話なんてわけの分かんないのがありまして。
Husband (H)	おーい、ばあさんや。
Wife (W)	はーいはいおじいさんなーんですか。
H	今あの、オ表を通ったのは、横丁のー源兵衛さんじゃあなかったかなあ。
W	ヨ言ってんですよおじいさん違いますよ今通ったのはあれ、横丁の源兵衛さんですよ。
H	アそうか、ハッハ、やああたしゃてっきり、横丁の源兵衛さんかと思ったよ。

A.1.8 #8 Ochi of “Jugemu”

From: Makura of 味噌豆 (*Misomame*)

Used in: Both Chapters 5 and 6

Note: “Jugemu” is a famous rakugo story.

Performer	ま寿限無というのはご存知だと思いますが長い名前を付けられた子が、あーガキ大将んなって近所の子供を棒でぶって泣かしちゃう。
——	泣かされた子が親御さんのところに言い付けに来るというのが、一番おしまいのとこで。
Kin-chan (K)	おばちゃーん。
——	おばちゃんとこのね、寿限無寿限無五劫の擦り切れ海砂利水魚の水行末雲行末風来末食う寝る処に住む処やぶらこうじのぶらこうじ。
——	パイポパイポパイポのシューリンガンシューリンガンのグーリンダイグーリンダイのポンポコピーのポンポコナーの長久命の長助が、おいらの頭アぶったから大きなたんこぶが出来たんだアア。
Jugemu’s mother (M)	アじゃ何かい？
——	金ちゃん、エエ？
——	えーウチの寿限無寿限無五劫の擦り切れ海砂利水魚の水行末雲行末風来末食う寝る処に住む処？

M やぶらこうじのぶらこうじパイポパイポパイポのシューリンガンシューリンガンのグーリンダイグーリンダイのポンポコピーのポンポコナーの長久命の長助が、金ちゃんの頭ぶってコブこしら、やだよ。

—– ちょっとお前さん聞いた？

—– ウチの寿限無寿限無五劫の擦り切れ海砂利水魚の水行末雲行末風来末食う寝る処に住む処？

—– やぶらこうじのぶらこうじパイポパイポパイポのシューリンガンシューリンガンのグーリンダイグーリンダイのポンポコピーのポンポコナーの長久命の長助が、金ちゃんの頭をぶってコブ拵えちゃたんだって。

Jugemu’s father 何をー？

—– じゃウチの寿限無寿限無五劫の擦り切れ海砂利水魚の水行末雲行末風来末食う寝る処に住む処やぶらこうじのぶらこうじパイポパイポパイポのシューリンガンシューリンガンのグーリンダイグーリンダイのポンポコピーのポンポコナーの長久命の長助が、金坊の頭アぶってコブ拵えしょうがねえなあの野郎は。

—– ああ金ちゃん悪かった。

—– 今な、あの薬付けてやるから頭出しなおお。

—– エーた。

—– ア、あれ？

—– 金ちゃん、コブなんかどこにもねえじゃねえか。

K ウェーン。

—– あんまり名前が長いからコブが引っ込んじゃった。

A.1.9 #9 Satsumaimo-taro

From: Makura of 元犬 (*Motoinu*)

Used in: Both Chapters 5 and 6

Note: Comic arrangement of Japanese famous folktale “Momotaro”

Performer (P)	エー昔々あるところに、おじいさんと、おばあさんが住んでおりましたある日、おじいさんが山へ柴刈りにおばあさんが、川へ洗濯に行きます。
—	アーおばあさんが川で洗濯をしていると川上から大きな、サツマイモが、ドンブラコードンブラコ流れてきました。
—	おばあさんが？
Elderly woman	おやなんて大きなお芋じゃろ、よし。
P	てんでウチい持って帰ると、おじいさんが帰ってこないうちにってんで焼き芋にして一人でこのサツマイモを食べちゃった。
—	トこんなにお腹が大きく張りまして、切なくなってー大きいのを一つブーっとやったら、このにおいがきついったってなんたって
—	風に乗ってフワフワフワーっと、山で仕事してるおじいさんの所までにおいが届いたもんですからおじいさんが柴を刈らずに、臭かったという。

A.1.10 #10 Hanasanka Jisan

From: Makura of 元犬 (*Motoinu*)

Used in: Both Chapters 5 and 6

Note: Comic arrangement of Japanese famous folktale “Hanasaka Jisan”

Performer (P)	エー犬のポチがある朝、エー庭でもってここ掘れワンワンここ掘れワンワンと、ええ仕事をして。
——	んでー庭に出たおじいさんがこれヒョイと見て。
Elderly man (E)	おやポチや、ああそんなところへ穴を掘ると人が転んで危ないよ？
——	やめなさーい。
P	それでもポチは、ここ掘れワンワンここ掘れワンワン。
E	ポチ、危ないから穴掘んのやめな。
P	それでもポチはここ掘れワンワンここ掘れワンワン。
——	いい加減むかっ腹が立ったおじいさん。
E	ポチ、穴掘んのやめないか。
P	てんで尻尾をギューっと引っ張った途端にポチが一言だけ口を利きました。
Dog Pochi	離さんかじいさん。

A.1.11 #11 Sagi No Ongaeshi

From: Makura of 元犬 (*Motoinu*)

Used in: Both Chapters 5 and 6

Note: Comic arrangement of Japanese famous folktale “Tsuru No Ongaeshi”

Performer (P)	エもう一つだけじゃあ、動物のお話のほうがなーんか評判がよさそうな気がするんで動物のお話を一つね。
——	えーっと、おじいさんが、雪の降る山道をエッチラオッチラ歩ってると、オー大きな、真っ白な、首の長い鳥が、罠にかかって、バタバタバタバタ苦しんでいる。
Elderly man (E)	おやおやかわいそうになあ、エエ？
P	エ心優しいおじいさん、ナ、罠からこの、オー白くて、首の長い大きな鳥外してあげると、嬉しそうに輪を書きながら振り返り振り返り山の向こうへと去って行きました。
——	ウチ帰って、日の暮れになりますと、表をトントン、トントン叩く者がある。
E	どなたじゃな？
P	ガラッと開けてみると真っ白な着物を着た、若い娘さんが。
E	お前さんは？
Young woman (Y)	はい、あたくしはおじいさんに大変に、お世話になりました。
——	御恩返しにあがりました。
E	恩返しとな。
Y	はい、あのーなんでございます。

Y ウ隣の部屋を貸してくださいましピタリと締め切って、決してどんな、物音がしても、音がしなくなるまで中を覗かないでくださいまし。

E ああそうか。

—— じゃあ、ウー隣の部屋へ入りなさい。

Y それではごめんあそばせ。

P 隣の部屋、ぴたりショオージを閉めるってえと、そのうちにトントンパターリ、トントンパターリという音がしてきた。

—— 何をしてるんだろうなおじいさん気になりましたが、音がしなくなるまでは覗かないという約束でございますからじっとしている。

—— ミヤガテなことにトントンパターリトントンパターリガタン、ガタン、トントンパターリトントンパターリ。

—— なんだろうな？

—— でも約束で覗かない。

—— トントンパタガタン、ガタン、ドシン。

—— ガタトントンパターリトントン何が起きてるでも気になってるけど、約束ですからヨが、シンと更けわたってもトントンパタズッシンズルズルガッタン、トントンパターリ。

—— なんだろうなあ気になるけど覗かない。

—— 明け方近くになってシーっと物音がしなくなったんでもうよかろう。

—— おじいさんがガラッ。

—— 障子を開けてみるってえと隣の部屋、家財道具が簞笥から何から一切合切なくなってた。

E ああ、あの鳥は鶴ではなくて、鷺だったんだ。

A.1.12 #12 Purachina Yatsume

From: The main part of お血脈 (*Okechimyaku*)

Used in: Chapter 6 only

Performer (P)	エー閻浮提金一寸八分ってえのはこんな小さなもんですが閻浮提金て何かって聞いたら、あの、白金の無垢だそうですな。
——	白金。
——	英語で言えば、プラチナですが。
——	まあこのープラチナはーまだしも白金という言葉が今若い子にピンとこないらしいですねえ。
——	ええーンー楽屋では、まあー落語家の卵、前座が、働いて、まあ落語家、ご存知の通り、エー身分、ンが今でもー分かれてますね下のほうから前座、二つ目、真打、ご臨終ってこういう順番に分かれてますが。
——	このーオー、ン前座ね、ええ。
——	ンもうウーあれですよ。
——	とにかくー師匠のウチでマインチ働いて寄席で、エーみんなにこき使われて、ええ。
——	上の者には絶対服従ですよ。
——	エ自由も人権もお金もなんにもないって、もう北朝鮮よりひどい暮らしをしてる。
——	エそれが前座です。
——	ね抑圧されてますからこの前座に。
Master (M)	おお、ウー前座。
Zenza (Z)	ンーなんなんなんすか。
M	ああ、お前白金知ってるか。
Z	へ？
M	白金だよ白金。
Z	ア、キャンドゥーですかダイソーですか。
M	百均じゃねえよ馬鹿野郎。

M 白金だよ。

Z は？

M だから、プラチナ。

Z ア、プラチナなら知ってますよ師匠。

—— あのこないだ、彼女にプレゼントした指輪がプラチナでしたから。

P てこれはムカッときましたよ。

—— 前座のくせに彼女がいて？

—— プラチナの指輪をプレゼントす思わず怒鳴っちゃいました。

M プラチなやつめ。

A.1.13 #13 Story About Mother Increasing Her Child’s Penalty

From: Makura of 真田小僧 (*Sanadakozo*)

Used in: Both Chapters 5 and 6

Performer (P)	アもう一杯ひどくなるってえと刑務所のそば行って懲役ごっこなんてんで、モッコ担いで土運んでる子どもがいまして。
Mother (M)	ちょいと。
— —	何してんだよお前はーエエ？
— —	ただいまじゃないの。
— —	着せて出してやりゃみーんな汚しつ帰っつくんだから。
— —	お前の洗濯でおっかさんイチンチなあんにも出きゃしないんだよ？
— —	何やってんのお前は。
Child (C)	アイしょうがないよんなガミガミ言ったってさ、あたい今みんなと懲役ごっこしてたんだから。
M	もっとまそもな遊びしなさいよおなあんだい懲役ごっこそれもいいけどね、エエ？
— —	オお隣の六ちゃんごらん六ちゃんを。
— —	お前と一緒に遊んでたって着物一つ汚しゃしないだろ？
— —	お前だって六ちゃんみたいに、綺麗におとなしく遊べないのかい？
C	おっかあなんにも知らねえんだから黙ってなよ。
— —	六ちゃんが汚れねえのにわけあんだよ。
— —	だって六ちゃん終身懲役だもん。

C もう生涯シャバの風に当たれねえ、仕方がねえからイチンチむしろの上へ座ってボーっとしてるしかねえって、可哀想なもんなんだいそこいくとさ、あたいはコソ泥で捕まったミツキの半端懲役だから、外役回されてモッコ担いで土運んだりなんかして、だから着物がすこーしぐらい汚れたってしゃあねえんだよ。

M たくしょうがないねえ。

—— じゃおっかさんがみんなに頼んであげるから、お前も明日から終身懲役にしてもらいな。

P なんてんで、親が子供の罪重くしたりなんかしまして。

A.2 味噌豆 (*Misomame*), the Test Story Used in Chapter 7

The transcription below is used for speech synthesis. It should be noted that detailed expression in the performance is different from performer to performer.

Danna (D)	定や、定吉。
Sadakichi (S)	へーい。
——	ダアサマお呼びでございますか。
D	あのな、台所でもって豆を煮てるんだ味噌豆を、うん。
——	煮えたかどうかちょいとお前見てきておくれ。
S	へーい。
——	ア、この鍋だなよいしょうわあーグツグツグツグツいってエーエー？
——	グツグツいってんの分かったけど、見ただけじゃ煮えたかどうか分かんねえや。
——	ああちょうどいいやしゃもじがあるから、一粒だけエ、よいしょ、あちい、あちい。
——	うまい。
——	よく煮えてるよ、フン、よく煮えてるの分かったけど、一つ食べるとあと引くなこりゃ、ああ。
——	よいしょ。
——	うん、うまい、うん。
——	うん。
——	うまい。
——	うん、うんうん？
——	ン？
——	ン？
——	うまい。
——	うん、うん。
——	うん、うまい、モ。

D　定吉。

S　ンー。

——　ア、旦那、あの、豆よく煮えてます。

D　煮えてますじゃないあたしは煮えてるかどうか見てこいった。

——　だーれが食べろと言いました。

——　じゃそこへ置いといて。

——　エーこのウお向こうの山田さんの所行ってな、この間頼んだ品物が、届いているかどうか、エ聞いてきておくれ。

——　いいか急いで行ってくるんだ分かったね。

——　まーったくしょうがないねあいつは。

——　ちょーいと目を離すってえとすぐにつまみ食いなんぞしてエエ？

——　まあ、うん、なるほど。

——　こらグツグツグツグツいってうまそうだよ。

——　エエ？

——　定吉は使いに行っちまって周りはだーれもいないし。

——　コにしゃもじがあるからじゃあたしも一粒だけ、エーよいしょ。

——　うん。

——　こりゃうまい。

——　ああよく煮えてるよ。

——　うーんよーく煮えてるの分かったけど、一つ食べるとあと引くねこりゃエエ？

——　よいしょ。

——　うん。

——　うんうん、うんこりゃうまい、うん。

——　うーん。

——　うーん、うん、待てよ？

——　こうやって食べてるところを定吉が帰ってきて見つかったらあれ、旦那だって食べてるじゃありませんかなんてんで、ええー立場がなくなっちゃうね、エエ？

D どうしようー見つかるのは嫌だけど豆は食べたい。
—— そうだ。
—— ここに茶椀があるから、エーこれへすこーしよそって、鍋の蓋をして、エーどこで食べようか二階ーはいけないね、隠れる場所がないよ？
—— 定吉がトントントーンと上がってきたらめっかっちゃうんだ。
—— 押入れのなーかーはー暗くって嫌だしねえ。
—— どーっか一人っきりんなれて誰にも見つからない所はイない、ないかな？
—— ええ一人っきりんなれて誰にも見つからないと。
—— あった。
—— 便所だよ。
—— 憚り憚り、ねえ？
—— あすこは一人入りゃ満員だからね、ええ。
—— ウー多少くさいけれどもな。
—— まあ我慢をして食べようじゃあないかここでいいや。
S 旦那ーただいま戻りましたー。
—— あのー山田さんお留守だったんで、あとでまた行ってきようと思うんでーすーけーどー。
—— 旦那ー。
S いないんですかー？
—— うわありがてえ、フ、旦那がいないあいだにねえ？
—— これこれこれうわあグツグツいってるよ、ハッハ。
—— うーん、ン、うまい。
—— うまい。
—— ア、あちゃちゃちゃちゃあっちゃア、ン、待てよ？
—— こうやって食べてるところをまたミ、旦那にめっかったらまた食べてんのかーって、今度は晩のおまんまア抜きんなっちゃうよ。
—— うん見つかるのは嫌だけど、豆は食べたいん。
—— エどう。
—— ああーこんなところに茶碗があるよねえ？

S この茶碗に、豆をいくらかよそって、で、蓋アして、さあどこで食べようかな。

—— 二階ーはいけないね、隠れる場所がないよ。

—— 旦那がトントントーンっと上がってきたらめっかっちゃう。

—— 押入れの中ーは暗くって嫌だしなあ。

—— どーっか一人っきりんなれて、誰にも見つからない所はないかな。

—— 一人っきりんなれて誰にもミ、あった。

—— 便所だよ。

—— エエ？

—— 憚り憚り。

—— あすこは一人入りゃ満員だからね、エーここでもってたーべーてーや。

—— ア、旦那。

D 定吉何しに来た。

S あのー、おかわり持ってきました。