

# **Analyzing and Predicting Prosodic Challenges for Japanese Students to Learn Oral English**

(英語音声学習における日本人学習者の  
韻律的課題の分析と予測)



程 禧璦  
**CHENG Xiai**

ID Number: 37-236545

Supervisor: Prof. Nobuaki Minematsu

Department of Electrical Engineering and Information Systems,  
Graduate School of Engineering,  
The University of Tokyo

*Master Thesis*

*Jan 2025*

## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

CHENG Xiai  
January 2025

## Acknowledgements

As time flies, in this ordinary yet extraordinary spring, I once again find myself approaching my graduation season.

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Nobuaki Minematsu and all lab members. Over the past two years, I have been constantly supported by their insightful guidance in every step of my research. I sincerely appreciate their patience in answering my questions and the careful review and revisions made to my thesis.

I am also profoundly grateful to my parents. Not only did they provide encouragement during moments of doubt and congratulate me on my achievements, but they have always respected and supported every decision I have made. Their unwavering support has been my strongest backing, and it is their silent dedication and encouragement that have given me the courage to face the unknown.

Finally, I would like to thank the University of Tokyo for providing such a pleasant environment for my two years of study. The university has given me a broader platform, allowing me to see a much larger world.

I wish all my teachers, family, and close friends the very best in all their endeavors, and I also look forward to the bright future ahead of me.

## **Abstract**

This study examines challenges Japanese learners face in learning oral English, focusing on prosody analysis from STEAC course data. Also, Generative Spoken Language Model (GSLM), sequence-to-sequence (seq-to-seq) Voice Conversion (VC), and Text-to-Speech (TTS) models were employed to analyze and predict student performance. Results show all models effectively forecast mimicry abilities, with proposed scenarios for optimal application, supporting tailored learning resources for diverse student needs.

# Table of contents

<b>List of Figures.....</b>	<b>vii</b>
<b>List of Tables .....</b>	<b>ix</b>
<b>1 Introduction.....</b>	<b>1</b>
1.1 Research Background .....	1
1.2 Research Objectives.....	2
1.3 Structure of This Thesis.....	2
<b>2 Related Work .....</b>	<b>3</b>
2.1 Challenges for Japanese in learning oral English.....	3
2.1.1 Vowel System.....	3
2.1.2 Consonant System .....	5
2.1.3 Pitch .....	7
2.1.4 Intonation.....	7
2.1.5 Speech Rhythm.....	8
2.2 Analysis of prosody control of the learners.....	9
<b>3 Prosody Analysis .....</b>	<b>11</b>
3.1 Introduction of STEAC .....	11
3.2 Tools Used for Analysis .....	14
3.2.1 Dynamic Time Warping .....	14
3.2.2 Forced Alignment.....	15
3.3 Data Processing .....	16
3.4 Prosody Analysis Of Overlapping Data in STEAC .....	16
3.4.1 Pitch control.....	17
3.4.2 Intensity contro .....	18
3.4.3 Duration control.....	21

<b>4 Modeling and Prediction .....</b>	<b>24</b>
4.1 GSLM .....	24
4.1.1 Introduction of GSLM.....	24
4.1.1.1 Baseline GSLM to simulate foreign accentuation .....	25
4.1.1.2 Modified GSLM to simulate foreign accentuation.....	26
4.1.2 Data Processing .....	27
4.2 Seq-to-Seq VC Model .....	27
4.2.1 Introduction of Seq-to-Seq VC Model .....	27
4.2.2 Data Processing .....	29
4.3 TTS .....	30
4.3.1 Introduction of ElevenLabs TTS .....	30
4.3.2 Data Processing .....	31
4.4 Experiment.....	32
4.4.1 GSLM models and their performance .....	33
4.4.1.1 Difficulty Estimation for Pitch Control .....	34
4.4.1.2 Difficulty Estimation for Intensity Control .....	35
4.4.1.3 Difficulty Estimation for Duration Control .....	37
4.4.2 VC/TTS models and their performance .....	37
4.4.2.1 Difficulty Estimation for Duration Control .....	40
4.4.2.2 Difficulty Estimation for Pitch and Intensity Control .....	41
4.5 Results and Discussion .....	42
<b>5 Conclusions and Future Works .....</b>	<b>44</b>
5.1 Conclusions.....	44
5.2 Future Work.....	44
<b>Reference .....</b>	<b>46</b>
<b>Appendix A   Tables.....</b>	<b>49</b>
<b>Appendix B   Publications .....</b>	<b>51</b>

## List of Figures

2.1 English Vowels and Distribution.....	3
2.2 Japanese Vowels and Distribution.....	3
2.3 The Japanese Constants .....	5
2.4 The English Constants .....	5
2.5 Consonant Articulation Positions .....	6
2.6 Graphic representation of the two language types .....	8
2.7 Visualization of the differences between learner and model audio...	10
3.1 course schedule for STEAC.....	12
3.2 Result of overlapping in STEAC.....	12
3.3 Result of good overlapping in STEAC.....	13
3.4 An example of the utility of dynamic time warping.....	14
3.5 An example warping path. ....	14
3.6 An example for FA .....	15
3.7 difference of pitch control.....	17
3.8 difference of pitch control.....	18
3.9 Comparison of intensity distribution between the native and learner audio.....	19
3.10 Comparison of intensity distribution between the model and learners .....	19
3.11 intensity difference between learner and native .....	20
3.12 intensity difference between learner and native .....	20
3.13 the result of duration control.....	21
3.14 the result of duration control.....	22
3.15 comparison of learner and model .....	22
4.1 Structure of GSLM .....	24
4.2 Architecture of the baseline GSLM.....	25
4.3 Architecture of the modified GSLM.....	26
4.4 Architecture of Seq-to-Seq VC.....	27
4.5 Architecture of Parallel WaveGAN.....	28
4.6 An example for generating audio use ElevenLabs TTS .....	31
4.7 Pitch Correlation Coefficients ((m)JP200 Learners) .....	35
4.8 Pitch Correlation Coefficients ((m)JP1000 Learners) .....	34
4.9 Intensity Correlation Coefficients ((m)JP200 Learners) .....	36
4.10 Intensity Correlation Coefficients ((m)JP1000 Learners) .....	36
4.11 Duration-based difficulty estimation in the modified GSLM .....	37
4.12 Pitch Correlation Coefficients for Week5 .....	38
4.13 Intensity Correlation Coefficients for Week5 .....	38
4.14 Duration Correlation Coefficients for Week5 .....	39

4.15 Pitch Correlation Coefficients for Week6 .....	39
4.16 Intensity Correlation Coefficients for Week6 .....	39
4.17 Duration Correlation Coefficients for Week6 .....	39
4.18 Pitch Correlation Coefficients for Week7 .....	39
4.19 Intensity Correlation Coefficients for Week7 .....	39
4.20 Duration Correlation Coefficients for Week7 .....	40



## List of Tables

3.1 correlation coefficients between students and native model audio ....	17
4.1 Division of the Training Dataset.....	30
4.2 Correlation coefficients between actual learners and generated audio (GSLM).....	33

# Chapter 1

## Introduction

### 1.1 Research Background

With the rise of globalization in the 21st century, English has become an essential global language for communication. As a result, more people around the world are eager to master English in order to engage in international communication. However, for non-native English speakers, particularly those without an immersive language environment, improving spoken English remains a significant challenge. This gap in language proficiency, especially in speaking skills, is a major issue in many educational systems. Traditionally, English education has focused on reading and writing skills due to limited resources, leaving speaking and listening skills underdeveloped. This situation highlights the need for more effective educational methods to improve English communication abilities, particularly in speaking and listening.

To address this challenge, the University of Tokyo launched the Special Training for English Academic Communication (STEAC) program, designed to enhance students' English communication skills, with a particular focus on listening and speaking. The STEAC program aims to optimize course structures by incorporating advanced techniques such as prosody analysis of learners' speech. This is crucial because Japanese and English differ significantly in rhythm, accentuation, and overall prosodic patterns, which means English spoken by Japanese learners often differs from that of native English speakers [1]. By understanding the unique prosodic features of English as spoken by Japanese learners, educators can better help students improve their pronunciation and intonation, making their spoken English more intelligible and closer to native-like speech.

One of the key aspects of improving spoken English through the STEAC program is determining the appropriate difficulty level for new learning materials. This is essential to apply the  $i+1$  principle [2], where each lesson is slightly more challenging than the previous one. Therefore, it is necessary to use a method that can accurately predict the difficulty level of new practice materials in terms of prosody control. Currently, this process is mainly carried out by experienced individuals. To reduce the workload, speech models are needed to generate sufficiently accurate prediction results. By continuously monitoring and adjusting the difficulty of practice materials and prosody imitation tasks, each learner is ensured to be continuously challenged in a supportive manner.

### 1.2 Research Objectives

The aim of this study is to provide the most suitable learning materials for Japanese native English learners. First, we reviewed and summarized the challenges faced by Japanese speakers in learning English, as identified in previous research. We then conducted a prosodic analysis (including pitch, intensity, and duration) on audio data collected from relevant English speaking courses to identify issues exhibited by the learners during their course participation.

To further improve the alignment of the course with the learners' proficiency levels, we applied GSLM, seq-to-seq VC models, and TTS models to model and predict the learners' speech. This allowed us to explore the feasibility of using these three models for predicting the difficulty of learning materials. Our analysis concluded that all three models are capable of predicting the difficulty of learning materials in different scenarios, providing technical support for the future development of courses.

### 1.3 Structure of This Thesis

This thesis is organized as follows. **Chapter 2** reviews existing research on Japanese learners' English speaking skills and highlights some common challenges faced by native Japanese speakers when learning spoken English. **Chapter 3** conducts a prosodic analysis of audio data collected from the STEAC course, identifying difficulties that Japanese speakers encounter in prosodic control—specifically in pitch, intensity, and duration—when overlapping with native English speakers' audio. **Chapter 4** utilizes three kinds of models: GSLM, VC, and TTS, to model and predict learners' performance. It also examines whether these models can effectively predict learning difficulties and discusses the applicable scenarios for each model. **Chapter 5** summarizes the entire thesis and provides suggestions for future research.

## Chapter 2

### Related Work

## 2.1 Challenges for Japanese in learning oral English

### 2.1.1 Vowel System

The vowel systems of Japanese and English differ significantly, presenting notable challenges for Japanese learners of English pronunciation (Fig. 2.1 and Fig. 2.2)[3][4]. Japanese has a relatively simple vowel inventory consisting of five distinct vowel sounds: /a/, /i/, /u/, /e/, and /o/. These vowels are pronounced clearly and consistently, with little variation in quality or length. In contrast, English has a much more complex vowel system that includes twelve distinct vowel sounds and multiple diphthongs. This greater variety introduces numerous pronunciation challenges for Japanese learners, who must navigate unfamiliar vowel contrasts and sounds absent in their native language.

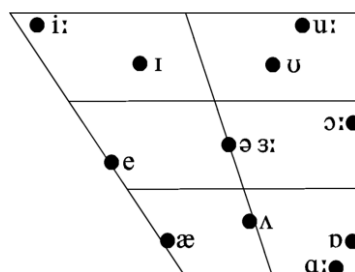


Fig. 2.1 English Vowels and Distribution

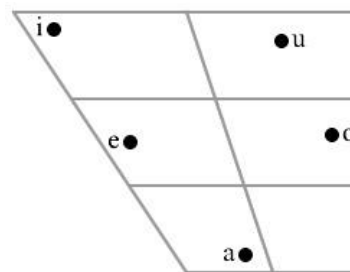


Fig. 2.2 Japanese Vowels and Distribution

One major difficulty for Japanese speakers is distinguishing between tense and lax vowels in English. For example, English differentiates between the tense vowel /i:/ in "see" and the lax vowel /ɪ/ in "sit." This contrast does not exist in Japanese, where vowel length and quality do not serve to differentiate word meanings in the same way [5]. As a result, Japanese learners may struggle to perceive and accurately produce these subtle distinctions, leading to mispronunciations that can affect intelligibility.

Another common challenge arises from the English schwa /ə/, an unstressed and neutral vowel sound frequently used in English, such as in the second syllable of "sofa." Since Japanese does not have a schwa sound, learners often replace it with more distinct and familiar vowels like /a/ or /o/. This substitution disrupts the natural rhythm and

stress patterns of English, making speech sound unnatural or overly emphasized. Misunderstanding vowel reduction, especially in unstressed syllables, is one of the most persistent pronunciation issues for Japanese learners.

These general difficulties in English vowel pronunciation can be categorized into four specific patterns observed among Japanese learners [5]:

### 1. Vowel Addition:

Japanese learners frequently add extra vowels after consonants due to their native language's phonotactic rules. In Japanese, syllables typically follow a consonant-vowel (CV) pattern, and words rarely end with a consonant (except for the nasal /n/). This structure encourages learners to insert vowels to make English words conform to familiar patterns. For example, the English word "smartphone" is often pronounced as /sma.to.fon/, where an /o/ vowel is added after the consonant sounds. This addition can affect pronunciation clarity and fluency, especially with English words that end in consonants or contain consonant clusters, and may also lead to an increase in the number of syllables in the vowels.

### 2. Vowel Extension:

Japanese learners may unintentionally extend vowel sounds, often due to influence from similar-sounding words in their native language. A notable example is the pronunciation of "baba" in "ali baba" as /ba:bə/. In Japanese, the word *bāba* with an extended vowel refers to an old woman or grandmother. This familiarity may lead learners to elongate the vowel in the English word, altering its intended pronunciation. Vowel extension can distort meaning and rhythm in English, particularly when learners apply Japanese vowel length rules to English words that do not require them.

### 3. Vowel Deletion:

Conversely, Japanese learners may sometimes omit vowel sounds, resulting in incomplete or fragmented word pronunciation. An example of this is the omission of the vowel /æ/ in the word "whereas" (/weər'æz/), which might be pronounced as /weərz/. This could occur due to unfamiliarity with the word or simply due to misreading, especially when the word appears at the beginning of a speech. Vowel deletion disrupts word structure and can make speech harder to understand, particularly when important syllables are dropped.

### 4. Vowel Substitution (Change of Vowel):

Japanese learners frequently substitute English vowels with those more familiar to them due to the limited vowel inventory in Japanese. For example, the English schwa /ə/, which does not exist in Japanese, is often replaced by /a/ or /æ/. In one case, students replaced the central schwa /ə/ with /æ/, a vowel positioned lower and more forward in the mouth. Since Japanese vowels typically avoid central tongue positioning—except for /a/ in a lower position—learners tend to favor mouth movements that feel more natural, leading to incorrect substitutions.

This pattern of vowel replacement can significantly alter word pronunciation and make speech sound less natural.

### 2.1.2 Consonant System

The consonant systems of English and Japanese differ significantly, contributing to pronunciation challenges for Japanese learners of English (Fig. 2.3 and Fig. 2.4) [5][7]. English has a much larger set of consonant sounds than Japanese, including several sounds that do not exist in Japanese [5]. Notably, Japanese lacks consonants produced in the labiodental (involving the lower lip and upper teeth), interdental (tongue between the teeth), and alveolar (tongue against the ridge behind the teeth) positions. These absent articulatory positions make it difficult for Japanese learners to perceive and produce certain English consonants accurately.

Fig. 2.3 The Japanese Constants

Place of Articulation		Bilabial	Alveolar	Alveopalatal	Velar	Glottal
Manner of Articulation						
Stops	Voiceless	p	t		k	
	Voiced	b	d		g	
Fricatives	Voiceless	Φ	s	Ç		h
	Voiced		z			
Nasals		m	n			
Liquids (Approximants)			r			

Fig. 2.4 The English Constants

Place of Articulation		Bilabial	Labiodental	Interdental	Alveolar	Alveopalatal	Velar	Glottal
Manner of Articulation								
Stops	Voiceless	p			T		k	
	Voiced	b			D		g	
Fricatives	Voiceless		f	θ	S	ʃ		h
	Voiced		v	ð	Z	ʒ		
Affricates	Voiceless					tʃ		
	Voiced					dʒ		
Nasals		m			N			
Retroflex Liquid					R			
Lateral Liquid					L			

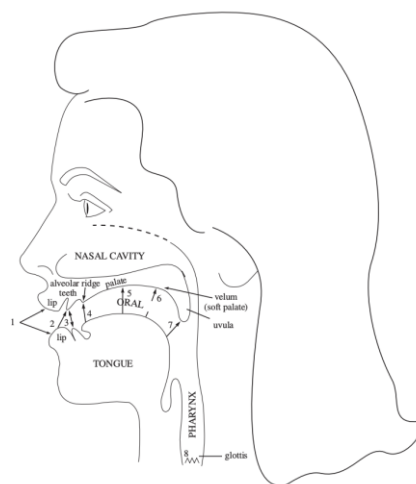


Fig. 2.5 Consonant Articulation Positions [36]

1. Bilabial 2. Labiodental 3. Interdental 4. Alveolar 5. Palatal 6. Velar  
7. Uvular 8. Glottal.

Japanese learners commonly exhibit three types of consonant-related pronunciation errors in English:

### 1. Change of Consonant [8]:

This was the most frequent type of error observed in learners' speech. Japanese learners often substitute unfamiliar English consonants with more familiar sounds from their native language. Key examples include:

- /θ/ → /s/: The voiceless interdental fricative /θ/ (as in "think") was commonly replaced by the voiceless alveolar fricative /s/ (as in "see").
- /ð/ → /d/: The voiced interdental fricative /ð/ (as in "this") was replaced by the voiced alveolar stop /d/ (as in "dog").
- /v/ → /b/: The voiced labiodental fricative /v/ (as in "very") was replaced by the voiced bilabial stop /b/ (as in "bat"), reflecting the absence of labiodental sounds in Japanese.
- /ɹ/ → /l/: The English /ɹ/ sound (as in "red") was often substituted with /l/, due to the ambiguous articulation of the Japanese /r/, which is a tap sound that lies between English /r/ and /l/.

And also in [8], it identified /f, v, θ, ð, w, l, ɹ/ as some of the most difficult English consonants for Japanese learners to master.

### 2. Consonant Addition:

Japanese learners sometimes insert extra consonant sounds when speaking English, often influenced by English spelling or their native phonotactic rules. For example:

The word "clothes" (/kloʊz/) was pronounced as /kloʊz.iz/, with an added /iz/ sound. This addition may result from interpreting the spelling of "clothes" literally, leading learners to overpronounce the plural suffix -es.

This pattern reflects how Japanese speakers may adjust English words to fit the typical syllable structure of Japanese, which often avoids consonant clusters and favors consonant-vowel patterns.

### 3. Consonant Deletion:

Another common issue is the omission of consonant sounds, especially word-final consonants, which are uncommon in Japanese. Specific examples include: dropping the final /r/ in words like "war", "counter", and "gangster".

Since Japanese words rarely end with consonants (except for the nasal /n/), learners often omit final consonants in English, particularly the English /r/, which does not exist in Japanese phonology [8].

### 2.1.3 Pitch

In English, pitch plays an essential role in marking word stress and conveying meaning. Stress patterns in English are often signaled by changes in pitch, loudness, and vowel length, which together highlight important words in a sentence. For example, in the sentence "I never said she stole my money," the word "never" would typically carry primary stress, and its pitch would rise accordingly. On the other hand, Japanese is a pitch-accent language, where pitch variations are used to distinguish word meanings, but these variations are more limited in scope compared to English [5][6][12].

In Japanese, the primary function of pitch is to distinguish between words or phrases, and the pitch movement within a word typically falls or rises in a relatively restricted range. Consequently, Japanese learners of English may over-rely on pitch for stress, leading to errors in English prosody. They may mistakenly use rising pitch in places where English requires a falling pitch, such as in declarative sentences, or they may misplace stress, which can affect the intelligibility of their speech.

### 2.1.4 Intonation

English has a highly nuanced intonation system, where variations in pitch indicate different sentence types (questions, statements, commands, etc.) and convey other nuances, such as emotions or emphasis. For example, rising intonation at the end of a sentence signals a yes/no question in English, while a falling intonation typically marks a declarative statement. In contrast, Japanese intonation is less variable and primarily serves to mark pitch accent on individual words [6][9]. The intonation patterns in Japanese are generally flatter than in English, and there are fewer pitch contours overall [5]. As a result, Japanese learners of English may encounter difficulty in producing the correct intonation patterns for different sentence types. They may produce English questions with rising intonation at the end, as they would in Japanese, even when falling



intonation is needed. Similarly, they may insert unnecessary pauses or breaks between words, particularly after adverbs or function words, reflecting the sentence-structuring tendencies in Japanese.

For example, in a sentence like "She completely forgot," Japanese learners might insert a pause after "completely," a pattern that is less common in English, where the adverb is usually attached to the verb without such a pause. This phenomenon is likely due to the learners' anticipation of a new sentence after adverbs, a structure more typical in Japanese. Additionally, learners may show a tendency to apply pitch variations that are typical in Japanese, such as raising the pitch in wh-questions, even when it deviates from English intonation patterns. These differences in prosody highlight the influence of L1 intonation patterns and the challenges faced by Japanese learners in mastering English intonation systems.

### 2.1.5 Speech Rhythm

In Japanese (syllable-timed language), stressed and unstressed syllables often receive similar emphasis in terms of syllable magnitude, reflecting the absence of strong and weak alternation in Japanese speech [6][10]. Unlike English, which is a stress-timed language where stressed syllables are typically longer, louder, and higher in pitch compared to unstressed syllables, Japanese is more evenly timed, with each syllable generally receiving a similar level of emphasis (Fig 2.5) [11]. This syllable-timed nature of Japanese means that there is less variation in syllable duration and pitch, leading to a more even-paced rhythm in speech. As a result, Japanese speakers may have difficulty perceiving and producing the appropriate stress patterns in English, where the rhythm is crucial for conveying meaning[6]. They might tend to place equal emphasis on syllables regardless of their stress status, leading to a flattening of the natural stress patterns found in English, which can affect the intelligibility and prosodic naturalness of their English speech. This difference in syllable emphasis is one of the key challenges for Japanese learners in acquiring English prosody.



Fig. 2.6 Graphic representation of the two language types

In summary, Japanese learners of English face several challenges due to the differences between the two languages. The limited vowel and consonant inventory in Japanese, along with the lack of consonant clusters, often leads to vowel substitutions and consonant changes in learners' English speech. In terms of pitch, Japanese relies on pitch accent, which contrasts with English's stress-based system, making it difficult for learners to master stress patterns. Additionally, intonation in Japanese is less varied, leading to issues such as incorrect rising intonation in English questions and unnatural pauses in sentences. The rhythm difference between the syllable-timed Japanese and stress-timed English further complicates learners' ability to produce the natural stress patterns required in English.

### 2.2 Analysis of prosody control of the learners

Shoda et.al show the prosody control results obtained in two experiments [13]. All the recordings were made by introducing to Japanese learners of English the task of expressive storytelling of picture books. Here, the learners have to imitate a model speaker's expressive storytellings by overlapping the learners' speech exactly on the model speech with their special attention paid to the intensity, pitch, and duration control made in the model speech.

In conducting prosodic analysis, the correlation coefficient between learner audio and native speaker model audio was used to assess prosodic control, and the differences between learner and native speaker audio were visualized (as shown in the fig. 2.7). The study revealed that Japanese English speakers often exhibit prosodic features such as inappropriate prominence with overly high pitch on unsuitable words, minimal stress variation, and shortened pronunciation in emphasized parts. Additionally, downward pitch shifts tend to show greater deviation than upward shifts.

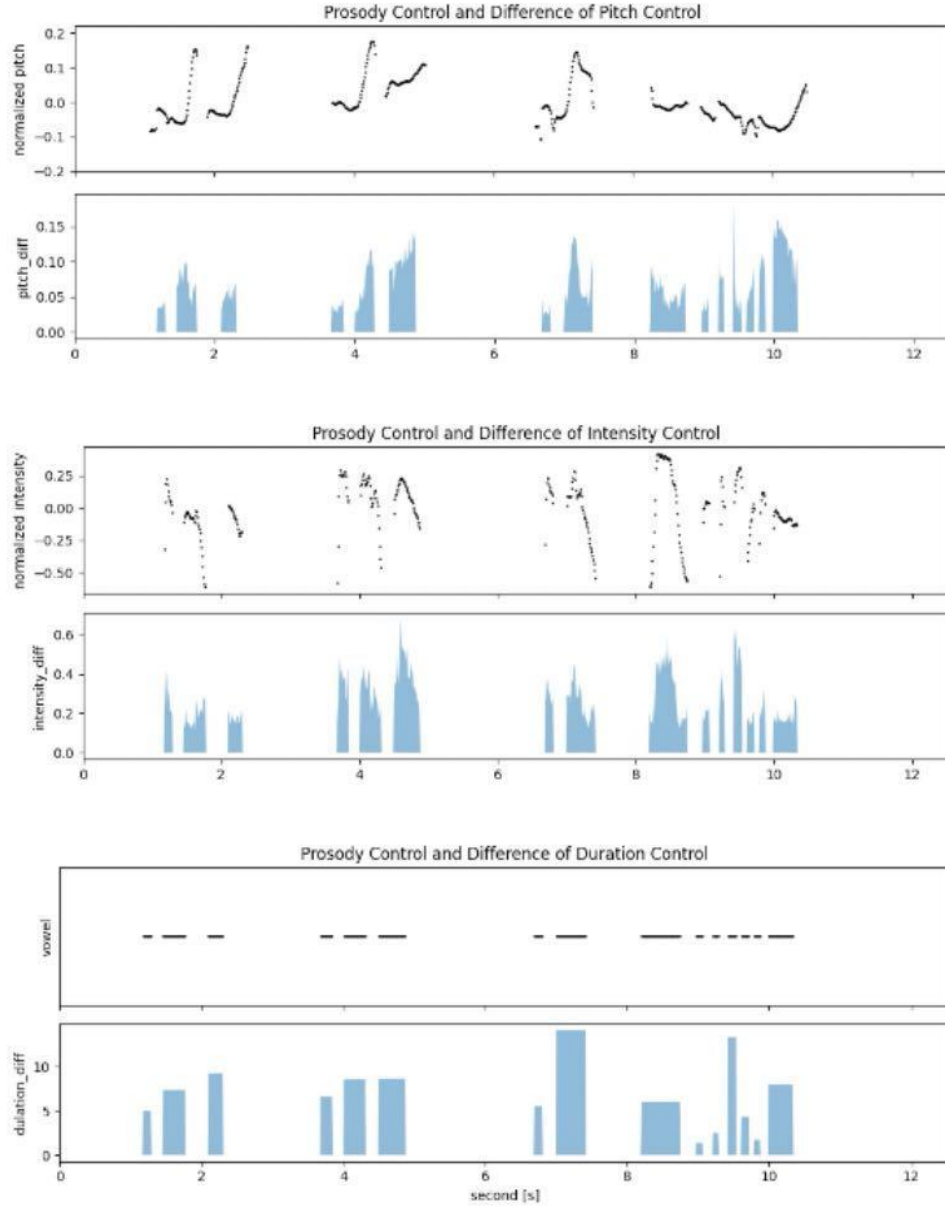


Fig. 2.7 Visualization of the differences between learner and model audio

This research employs the same prosodic analysis method to analyze the data from STEAC 2023S data.

## **Chapter 3**

### **Prosody Analysis**

This chapter mainly introduces the STEAC course and presents a prosodic analysis of learner audio collected from STEAC 2023S. It discusses the challenges native Japanese speakers face in prosodic control when learning spoken English.

#### **3.1 Introduction of STEAC**

In this study, we collected the analysis data from STEAC [14]. STEAC (Special Training for English Academic Communication) is an original course developed and offered by Professor Nobuaki Minematsu of the Department of Engineering at the University of Tokyo. The objective of this course is to enhance students' English listening and speaking skills, particularly targeting those who have limited exposure to an English-speaking environment. The course aims to develop an ear for understanding various English accents and distorted audio caused by noise and channel distortions, not only focusing on North American English but also encompassing global variations. Additionally, it seeks to improve students' ability to express themselves effectively in English by explaining the differences in audio patterns between Japanese and English, starting with rhythm, progressing to intonation, and delving into the pronunciation of individual sounds. Leveraging widely-known speech technologies, such as smart speakers, the course employs audio analysis to evaluate students' English speech and provide constructive feedback. Furthermore, daily assignments spanning August and September are designed to acclimate students to listening to and speaking English regularly, fostering a continuous learning environment.

2023 8 August 令和5年 葉月							2023 9 September 令和5年 長月						
MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	SUNDAY
instruction	x0.8 shadowing 30-sec speech extracted for textbook					6	instruction	Expressive speech shadowing Natural and synthetic voices are used.					3
instruction	+ listening comprehension test + prosodic overlapping				2	13	instruction	Noisy or distorted speech shadowing Re-synthesized voices are used.					10
instruction	x0.8 → x0.9 + ASR (word-based pron. deviation)				9	20	instruction	World Englishes speech shadowing Natural and synthetic voices are used.			5		17
instruction	x0.9 → x1.0 + phoneme-based pron. deviation				5	27	instruction	TED talk shadowing Commenting while viewing videos			3	休みの日	24
28	29	30	31	1	2	3	instruction	Rephrasing Summarizing presentation videos					1

Fig. 3.1 course schedule for STEAC

The STEAC 2023S takes place during the summer vacation, spanning from July 31 to October 1, a total of nine weeks (63 days), with approximately 30 minutes dedicated to each session per day. Fig. 3.1 shows the course schedule for STEAC 2023S. STEAC 2023S comprises six tasks: shadowing, overlapping, script shadowing, reading, rephrase, and describing. This analysis primarily focuses on the task type of Overlapping. Overlapping involves repeating the prompted audio in a manner where the learner's speech overlaps with the provided audio. Immediately after vocalization, the rhythm and intonation overlays are visualized and scored, resembling a karaoke game.

The figure below (Fig.3.2) depicts the learner's interface during the recording task in STEAC. The screen displays a visual native model audio and comparison between the learner's audio and the model's audio, with the model's audio represented in black and the learner's audio in green. The upper portion shows the syllable magnitude, while the lower portion illustrates the pitch comparison. A higher degree of overlap indicates a better imitatio effect (Fig. 3.3 represents a result with improved imitation). On the



Fig. 3.2 Result of overlapping in STEAC

right side, values for syllable magnitude and pitch of both the learner's and model's audio are computed.

Syllable magnitude in the context of comparing learner's audio and model's audio refers to the measurement of the strength or intensity of syllables in the spoken words.

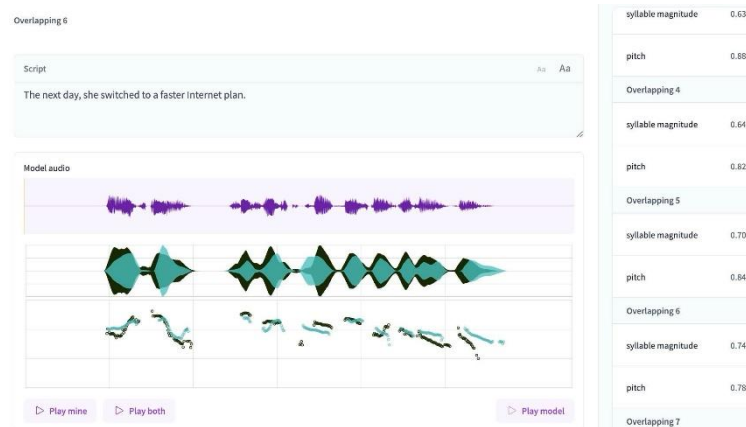


Fig. 3.3 Result of good overlapping in STEAC

It reflects how distinctly the syllables are pronounced in both the learner's and model's utterances. A higher syllable magnitude suggests clearer and more emphatic pronunciation of syllables. Pitch refers to the frequency of a sound, and it is generally associated with the perceived highness or lowness of the sound. Specifically, pitch reflects the fundamental frequency of the sound, which is the speed of the sound vibrations. When comparing learner's audio and model's audio, analyzing pitch provides information about the differences in pitch between the learner's pronunciation and the model's pronunciation. A successful imitation often involves the learner accurately reproducing the pitch variations and contours present in the model's audio.

From STEAC 2023S, the following samples were extracted and used for analysis. All the samples were recorded at prosodic overlapping practices. A passage is about 30-second long, which are composed of 5 to 7 phrases. Prosodic overlapping was conducted phrase by phrase. A total of 11,391 phrase recordings were used in this study.

- Week2 Day3 to Day7 at x0.8 speaking rate
- Week3 Day2 to Day7 at x0.9 speaking rate
- Week4 Day2 to Day7 at x1.0 speaking rate
- Week5 Day2 to Day7 at x1.0 speaking rate
- Week6 Day2 to Day7 at x1.0 speaking rate
- Week7 Day2 to Day7 at x1.0 speaking rate

## 3.2 Tools Used for Analysis

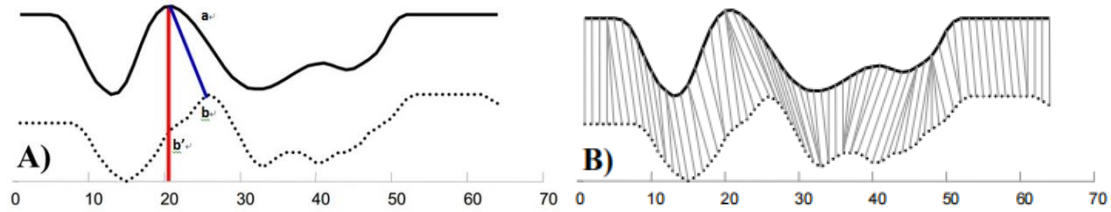


Fig. 3.4 An example of the utility of dynamic time warping.

### 3.2.1 Dynamic Time Warping

Dynamic Time Warping (DTW) [15] is a powerful algorithm extensively used in speech processing and time-series analysis, particularly when comparing sequences that vary in timing, speed, or length. In speech processing, DTW helps align sequences with similar content but different speeds, such as spoken words or phrases. For example, as shown in Figure A, the same sequence is recorded on different days. Although the sequences share similar shapes, they are not aligned temporally, and using a simple distance measure that pairs corresponding points would result in an overly pessimistic dissimilarity. However, DTW, demonstrated in Figure B, aligns the sequences

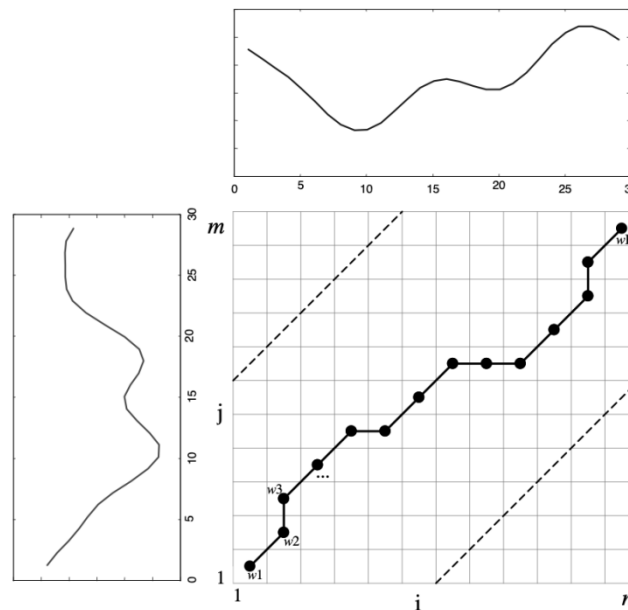


Fig 3.5 An example warping path.

effectively, allowing for the calculation of a more refined and accurate distance measure.

The DTW algorithm computes the optimal alignment between two sequences by minimizing the cumulative distance between corresponding points, allowing for stretching or compressing of the sequences to align them, even when they differ in timing. A distance matrix is used to represent the differences between each point in the sequences, and the algorithm finds a warping path (in Fig. 3.5  $w_1, w_2, \dots, w_k$ ) through this matrix that minimizes the total distance. This path adjusts for timing variations across the sequences.

By applying Dynamic Time Warping (DTW), a file is generated that shows the correspondence between each frame of the audio data from A (native model audio) and B (learner audio).

#### 3.2.2 Forced Alignment

Forced Alignment (FA) aligns audio with its phonetic or word transcription by first generating the probability of each phoneme label for every audio frame, forming an emission matrix. Using this matrix, a trellis is constructed to map the likelihood of each phoneme occurring at each time frame. The algorithm then finds the most probable

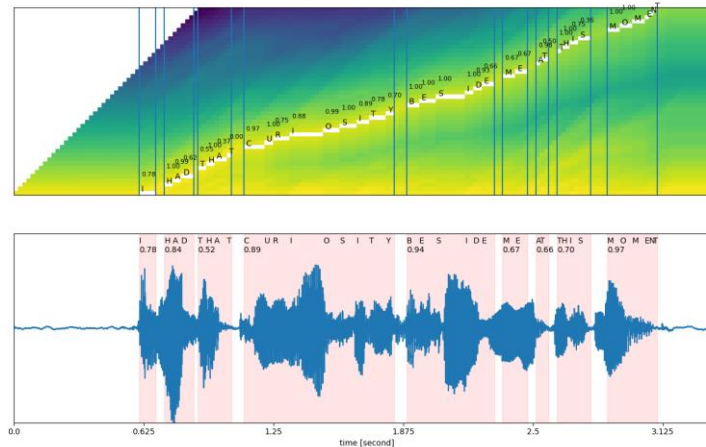


Fig 3.6 An example for FA

alignment path through the trellis, resulting in precise timing information that links each audio frame to its corresponding phoneme.

By performing FA processing, it becomes possible to identify which phoneme was produced in each frame. Using the results output by the FA processing, the start and end positions of vowel segments with clear control over pitch and intensity were determined. In cases where multiple pronunciations are allowed for a word, different phoneme sequences may be output between the model voice and the learner's voice. Therefore,



the FA processing was only performed on the model voice, and by comparing the results of the FA processing with the DTW results, the vowel segments in the learner's voice were identified.

## 3.3 Data Processing

Previous studies have shown that vowels play a significant role in prosody control [16]. Therefore, this study focuses specifically on the prosodic control of vowels. For each model speech sample and its corresponding transcript, forced alignment was performed to automatically detect the starting and ending time indices of all vowels. Following this, both the model and learner audio samples were converted into Phonetic PosteriorGrams (PPGs), PPGs are a form of phoneme-based speech feature representation. PPGs provide a posterior probability distribution of different phonemes for each time frame, creating a time-phoneme matrix from the input speech signal. This method captures the phone-level details in the speech signal and it is speaker-independent. After that, DTW [15] and FA was applied to time-align each learner sample with its corresponding native model sample. Vowels in the learners' audio were then automatically identified using the method described in [16].

The WSJ-KALDI [17] recipe was utilized for PPG conversion, and DTW was executed with symmetric local path constraints [18] to ensure accurate alignment. Additionally, the Short-Time Fourier Transform (STFT)[19] and the Speech Signal Processing Toolkit (SPTK) [20] were used to extract pitch and intensity information from the model audio along the time axis. This process enabled the numerical representation of prosodic features.

Subsequently, for each aligned vowel, the prosodic features—pitch, intensity, and duration—were extracted from both the learners' and native speakers' model audio. The correlation coefficients of these features were then calculated. These correlation coefficients were used as indicators of imitation quality: higher correlation values suggest that the learner's prosodic features are more similar to those of the native speaker, indicating better mastery of the corresponding prosodic feature.

## 3.4 Prosody Analysis Of Overlapping Data in STEAC

The weekly calculation results of the correlation coefficients between learners' audio and the native model audio for pitch, intensity, and duration are as follows (Table 3.1).

It can be observed that pitch imitation performance was the highest in the first week. However, as the speech rate increased over time—progressing from Week 2 (0.8×) to Week 3 (0.9×) and then to Week 4 (1.0×)—a noticeable decline in pitch imitation was observed in the second week. This was followed by a slight improvement in the third

Table 3.1 correlation coefficients between students and native model audio

	Pitch	Intensity	Duration
Week2 (0.8x)	0.731	0.536	0.719
Week3 (0.9x)	0.701	0.551	0.749
Week4 (1.0x)	0.709	0.559	0.740

week, likely due to learners adapting to the faster speech rate after an additional week of practice.

In contrast, intensity imitation showed continuous improvement across all weeks. This steady progress suggests that learners gradually became more adept at controlling intensity, regardless of the increasing speech rate.

For duration imitation, performance improved from Week 2 (0.8x) to Week 3 (0.9x) but began to decline by Week 4 (1.0x). This pattern indicates that while moderate increases in speech rate may initially enhance learners' control over duration, further acceleration can hinder their ability to maintain accurate timing.

Below is a separate analysis of the pitch, intensity and duration of the audio, and some analysis results are explained for each.

#### 3.4.1 Pitch control

In this section, the pitch values extracted from the native model audio were first normalized. This normalization involved taking the logarithm of the pitch values and then subtracting the overall mean. The standardized pitch values were then visualized in the upper graph to illustrate the pitch variation trends of the native speaker's speech.

Next, the difference between the standardized pitch values of the native speaker's audio and the average of learner's audio was calculated. This difference was plotted in the lower graph to highlight areas where the learner's pitch imitation deviated from the native speaker's model audio. In this graph, the blue line represents the magnitude of the difference: the higher the value, the greater the discrepancy between the learner's

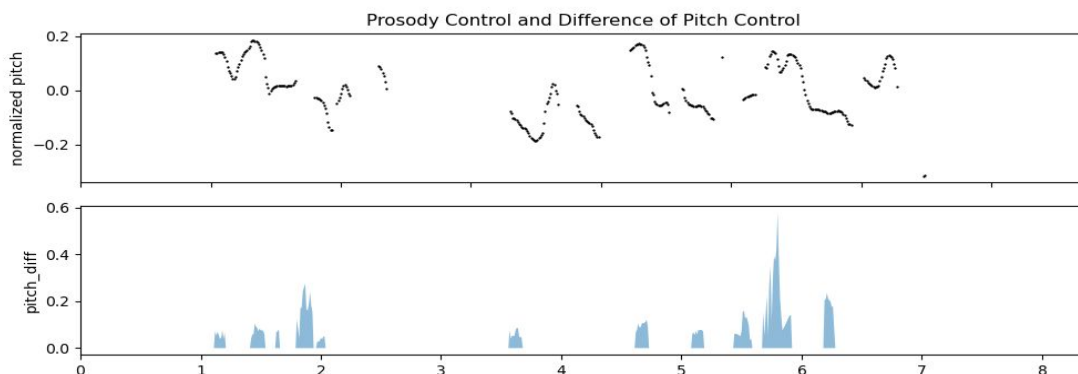


Fig. 3.7 difference of pitch control

*His mother is a chef and Richard's sandwiches are **always** delicious.*

「Richard の昼食(W2D4)」

and the native speaker's pitch. Therefore, larger values on the blue line indicate poorer imitation performance by the learner at those points in the speech.

The largest difference observed in the graph (Fig. 3.7) occurs during moments of rapid pitch change. Specifically, the word "ALWAYS" (around 6.0s) exhibits a noticeable pitch movement, where the pitch first decreases and then sharply increases.

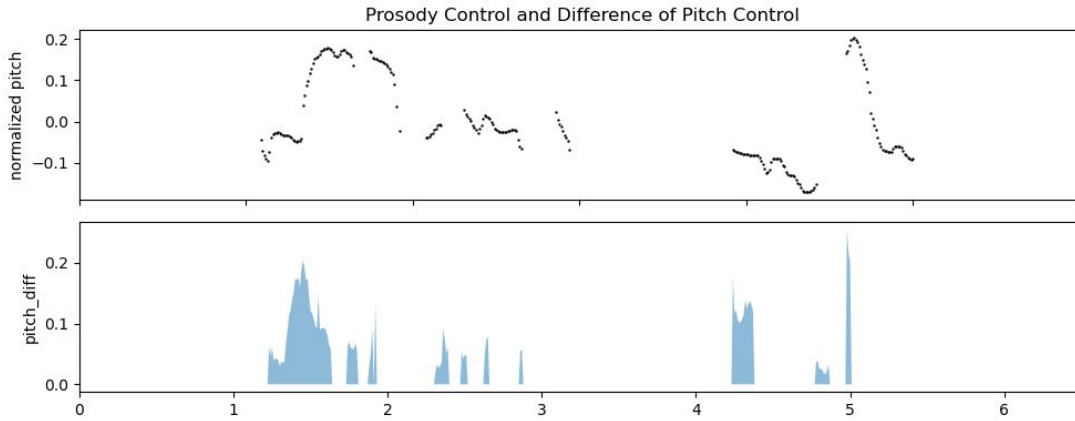


Fig. 3.8 difference of pitch control

*Veronica happily accepted and in return*  
 「Richard の昼食(W2D4)」

A similar situation can be observed in Fig. 3.8, where rapid pitch changes occur in the words **VERONICA** (around the 1st second) and **RETURN** (around the 5th second). Both of these words exhibit sharp fluctuations in pitch, similar to what was seen with the word "ALWAYS" in Fig. 3.7. These instances further illustrate the challenges learners face when attempting to imitate speech with significant pitch variations.

Based on the comprehensive analysis above, we can conclude that learners struggle more with imitating sections of speech that involve substantial pitch changes. This finding is consistent with the results reported in reference [13], which also highlighted the difficulties learners encounter in replicating rapid pitch transitions.

#### 3.4.2 Intensity control

In this section, the intensity values were first converted into numerical form and then standardized. The standardized results were subsequently visualized in the graph (Fig. 3.9). In the graph, the black dots represent the intensity distribution of the native speaker's audio, while the blue dots indicate the intensity distribution of the learner's audio. Here, the intensity distribution of two phrases from the same specific learner during a day's task is displayed in the graphs (Fig. 3.9 and Fig.3.10).

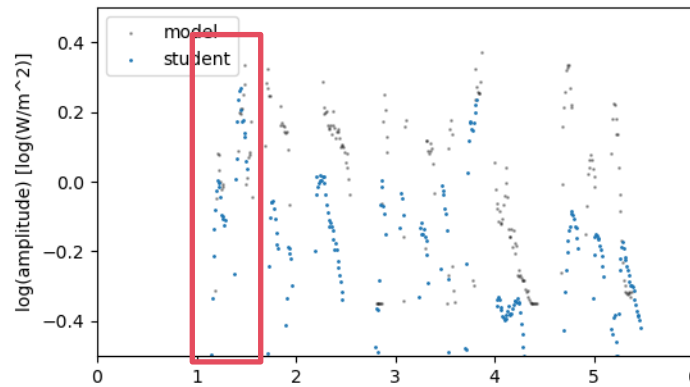


Fig. 3.9 Comparison of intensity distribution between the native and learner audio

*She bought him a snack at the cafeteria for dessert.*  
 「Richard の昼食(W2D4)」

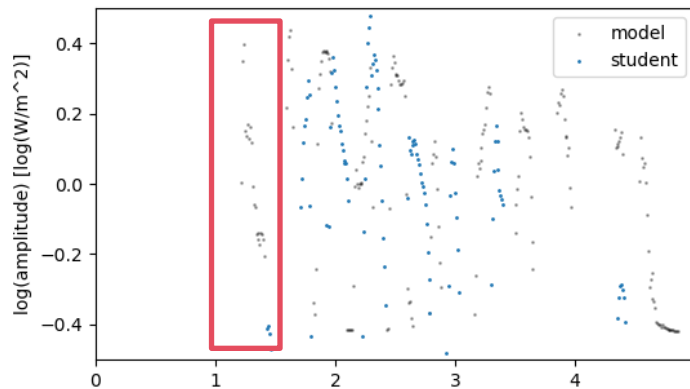


Fig. 3.10 Comparison of intensity distribution between the model and learners

*so Richard offered to share his lunch with her.*  
 「Richard の昼食(W2D4)」

By comparing Fig. 3.9 and Fig.3.10, we can observe that in Fig. 3.9 the imitation of the intensity at the beginning of the audio is relatively weak, but the overall imitation effect improves as the task progresses. In contrast, Fig.3.10 shows that when the model audio has a higher intensity at the beginning, the imitation effect tends to be worse. This suggests that learners may struggle more with imitating high-intensity audio at the start of the task, possibly due to the difficulty in mimicking the intensity at the outset.

### 3. Prosody Analysis

Next, the difference in intensity control between the native speaker's audio and the learner's audio is displayed using the same method as in the previous section. Similarly, the blue line in these graphs represents the magnitude of the difference: the higher the blue line, the larger the discrepancy, which indicates a poorer imitation effect.

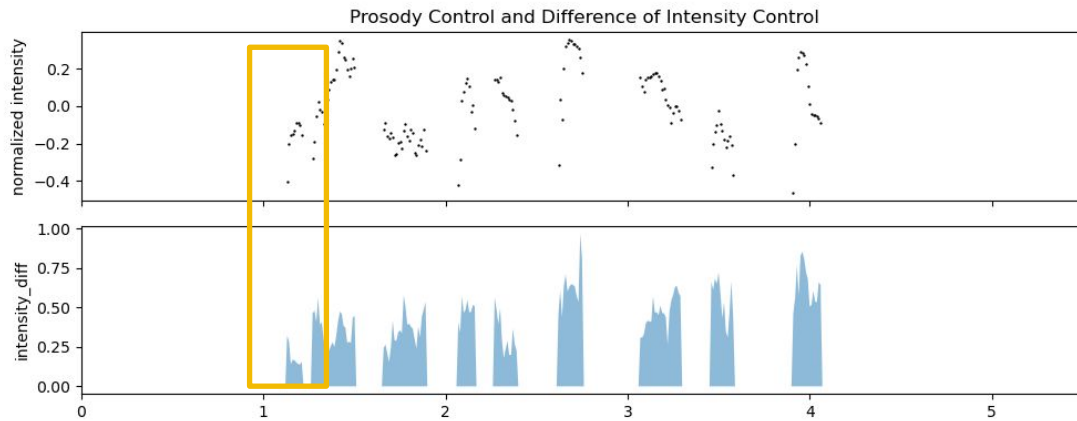


Fig. 3.11 intensity difference between learner and native

*But only if she got good grades at school.*  
「Sierra とビデオゲーム(W2D4)」

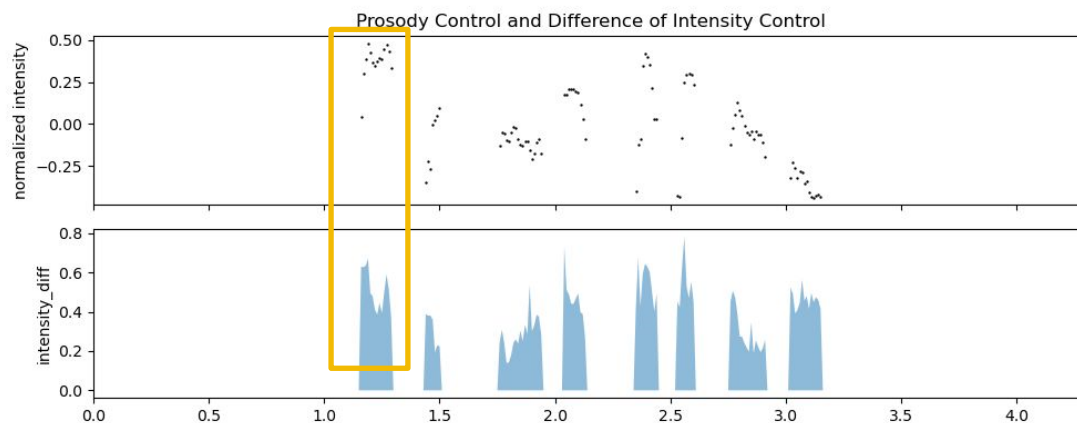


Fig. 3.12 intensity difference between learner and native

*They told her could have it,*  
「Sierra とビデオゲーム(W2D4)」

From these figures (Fig. 3.11 and Fig. 3.12), it is evident that when the model's audio begins with a higher intensity (as highlighted in the yellow box), learners find it significantly more challenging to replicate the intensity accurately. The higher initial intensity appears to create a larger gap between the learner's imitation and the model, making it more difficult for the learner to maintain the same level of intensity.

### 3.4.3 Duration control

For the analysis of duration control, the difference in vowel length between the native speaker's audio and the learner's audio was visualized in the same way as previous sections, as shown in Fig. 3.13. In the upper graph, the black line represents the duration of each vowel in the native speaker's audio. The lower graph shows the difference in duration between each vowel in the learner's audio and the corresponding vowel in the native speaker's audio.

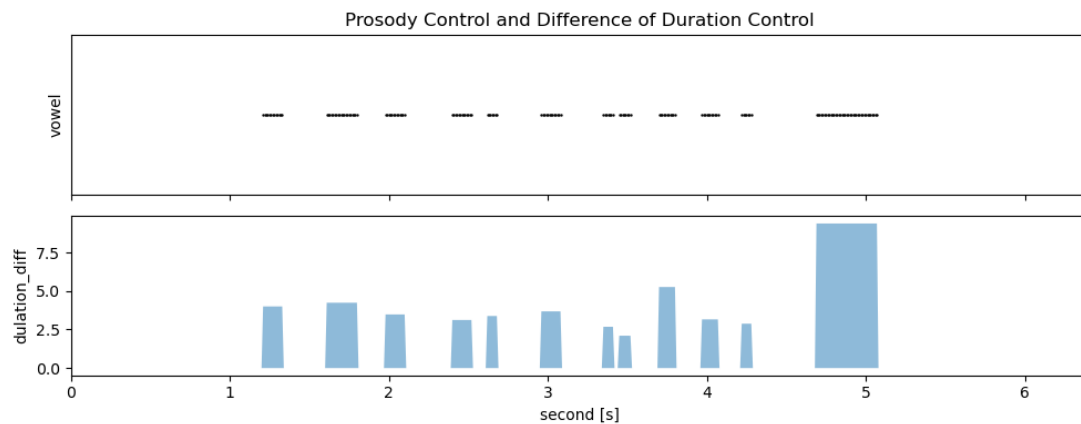


Fig. 3.13 the result of duration control

*so Steve thinks he can start giving her lessons soon.*  
 「Steve 親子のプール教室(W2D7)」

In this visualization, the blue boxes indicate areas where the difference in vowel duration is larger. The higher the value within the blue box, the greater the discrepancy between the learner's vowel duration and the native speaker's vowel duration.

From the image (Fig. 3.13), it can be observed that there is a significant discrepancy between the learner data and the model data when pronouncing the long vowel in the final word, "soon." This suggests that when the model's vowel duration is longer, learners tend to exhibit larger differences in their attempts to imitate the vowel duration accurately. This finding aligns with the results presented in previous

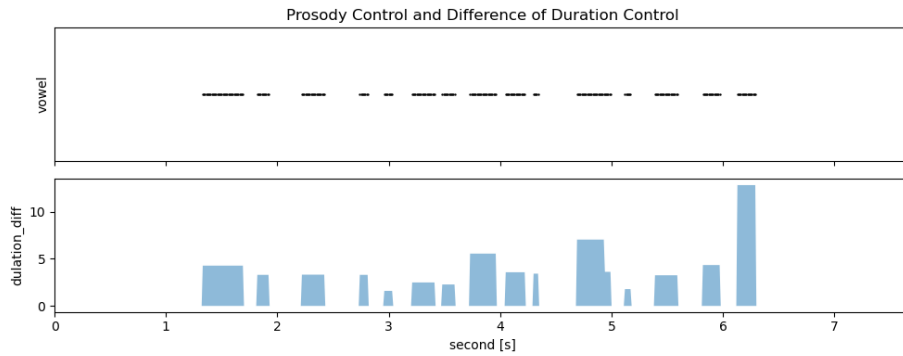


Fig. 3.14 the result of duration control

*Kenta and his friends have decided that they will stay in the same*  
*hotel.*  
 「Kenta と友人の旅 行計画(W2D7)」

studies, such as those in papers [13] and [20], which also concluded that longer vowel durations pose greater challenges for learners in terms of accurate imitation.

There are also some different conclusions, as seen in Fig. 3.14. The learner data shows significant differences, particularly at the final vowel "e", despite its relatively short duration in the model's audio. This suggests that learners might be overemphasizing or prolonging certain vowel sounds.

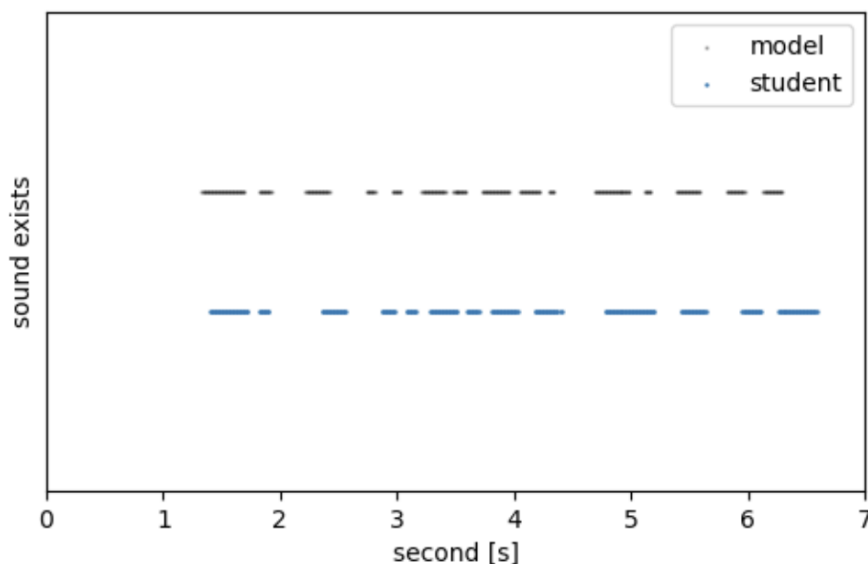


Fig. 3.15 comparison of learner and model

*Kenta and his friends have decided that they will stay in the same*  
*hotel.*  
 「Kenta と友人の旅 行計画(W2D7)」

Fig. 3.15 further illustrates this by comparing the duration of each vowel segment between the model and the learners' data. The longer the line in the graph, the longer the duration of that particular segment. From this comparison, we can see a notable extension in the final vowel segment in the learners' data, particularly for the vowel "e".

This behavior could be related to the learners' pronunciation habits, such as those seen with the Japanese word ホテル(hotel), where learners might be accustomed to prolonging the final "e" sound. These phenomena, such as vowel extension and the Katakana effect, are also discussed in reference [21], which points to how these specific habits influence vowel duration during imitation.



## Chapter 4

### Modeling and Prediction

The previous chapter explored the topics that are particularly challenging for Japanese learners to mimic in English speech. In order to arrange lessons that align with the “i+1” principle, it is necessary to rank new information materials by their difficulty level. In this chapter, three models—GSLM, VC, and TTS—are primarily utilized to model learners' audio. In the GSLM model, only the native speaker's audio is required as input, and there is no need to use learner's audio for training. The VC model, on the other hand, requires paired input of both the native speaker's audio and the learner's audio for training. The TTS model requires the learner's audio as input, along with the native speaker's audio, which needs to be provided in text form for prediction. It examines the similarity between the generated audio and the learners' authentic audio, as well as the feasibility of using these models for predicting difficulty in different scenarios.

#### 4.1 GSLM

##### 4.1.1 Introduction of GSLM

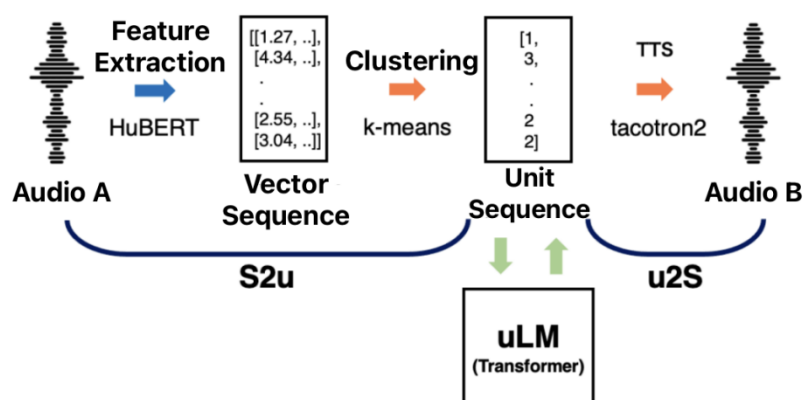


Fig. 4.1 Structure of GSLM

The Generative Spoken Language Model (GSLM) [22] is designed to replicate humans' natural ability to speak without relying on written text comprehension. It leverages *units*, which are pseudo-text representations generated through self-

supervised learning, allowing spoken language to be processed similarly to traditional text.

GSLM operates by integrating several key components:

1. **Speech-to-Unit (S2u):** This module converts raw audio into discrete units using models like wav2vec 2.0 [23] or HuBERT [24], both of which excel in self-supervised learning for speech representation.
2. **Unit-based Language Model (uLM):** Built with Transformer architecture, this model processes the unit sequences, enabling "textless NLP" where speech is modeled without textual data.
3. **Unit-to-Speech (u2S):** This component reconstructs speech from unit sequences using models such as Tacotron2 [25], allowing natural speech generation from non-textual representations.

By combining these modules, GSLM demonstrates that unit-based models can achieve performance comparable to traditional Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) pipelines. This approach highlights the potential of GSLM in advancing speech generation and processing without relying on conventional text-based methods.

### 4.1.1.1 Baseline GSLM to simulate foreign accentuation

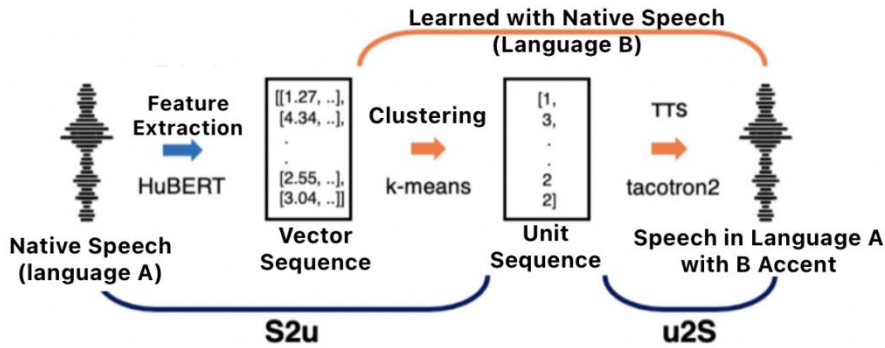


Fig. 4.2 Architecture of the baseline GSLM

Using the improved Generative Spoken Language Model (GSLM) (Fig. 4.2), the process of foreign accent formation can be simulated to synthesize speech with authentic foreign accents[26]. This is achieved by training the Speech-to-Unit (S2u) clustering model and the Unit-to-Speech (u2S) model in GSLM with speech corpora from native speakers of a target language (denoted as Language A in the diagram), which differs from the input language (Language B).

In this setup, the S2u model learns to encode the phonetic and prosodic characteristics of Language A, while the u2S model is trained to reconstruct speech using these learned patterns. When speech from Language B is processed through this system, the output speech naturally incorporates the accentual features of Language A,

effectively simulating how a native speaker of Language A might speak Language B with a foreign accent.

This capability highlights the model's potential for nuanced speech synthesis, enabling more accurate simulations of cross-linguistic accent perception and production.

#### 4.1.1.2 Modified GSLM to simulate foreign accentuation

In the baseline GSLM, the Unit-to-Speech (u2S) module directly processes discrete token sequences as input, preserving the input's temporal structure. However, this design limits the model's ability to realistically simulate duration-based accent features, as it cannot effectively adjust speech rhythm or timing to reflect natural accent variations.

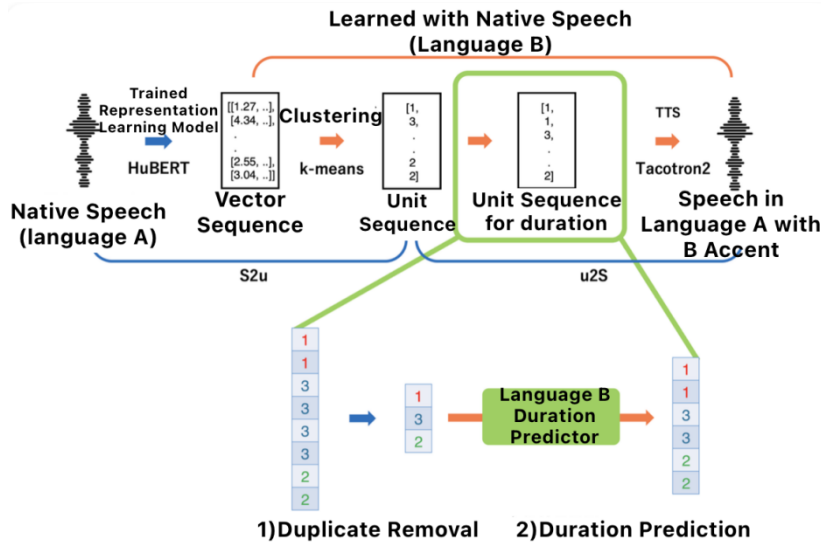


Fig. 4.3 Architecture of the modified GSLM

To address this limitation, a duration adjustment module was integrated into the original GSLM architecture before the u2S component [27] (Fig. 4.3). This enhancement involves two key steps:

1. **Duplicate Token Removal**: Redundant consecutive tokens are removed to eliminate unnecessary repetition, streamlining the input sequence and reducing distortions in speech rhythm.
2. **Duration Prediction**: A predictor based on FastSpeech2 [28] is employed to generate phoneme durations that align with the target accent's natural speech patterns. This model allows dynamic control over speech timing, effectively adjusting the prosody to match the intended accent.

By incorporating this duration adjustment module, the improved GSLM can produce synthesized speech that more accurately reflects the natural timing, rhythm, and accentual characteristics of the target language (Language B in fig.4.3) [26]. This leads to more realistic and expressive speech generation, enhancing the simulation of foreign accents.

### 4.1.2 Data Processing

Previous research in [26] demonstrated that the number of units used in the GSLM significantly affects the performance of foreign accentuation. Specifically, varying the number of units led to differences in how accurately the model could replicate foreign-accented speech. Additionally, findings from [22] indicated that increasing the number of units resulted in more natural-sounding speech in analysis-resynthesis tasks, highlighting the importance of unit granularity in speech quality.

Building on these insights, the current study employed the original GSLM model for foreign accentuation [26] to generate Japanese-accented speech using three different unit configurations: 50 units (JP50), 200 units (JP200), and 1000 units (JP1000). Furthermore, the modified GSLM model—enhanced with a duration adjustment module—was also used to synthesize speech with the same unit configurations: 50 units (mJP50), 200 units (mJP200), and 1000 units (mJP1000). This comparison between the original and modified models across different unit sizes aimed to evaluate the similarity between the synthesized speech and the learners’ real audio. The goal was to determine which model could generate more prosodically realistic speech and identify the most suitable model for predicting the difficulty of prosody imitation.

Since the GSLM model does not require any learner data as input, it can be used to make a general difficulty prediction for all practice materials before the course begins. This allows for the determination of the course plan for the first three weeks.

## 4.2 Seq-to-Seq VC Model

### 4.2.1 Introduction of Seq-to-Seq VC Model

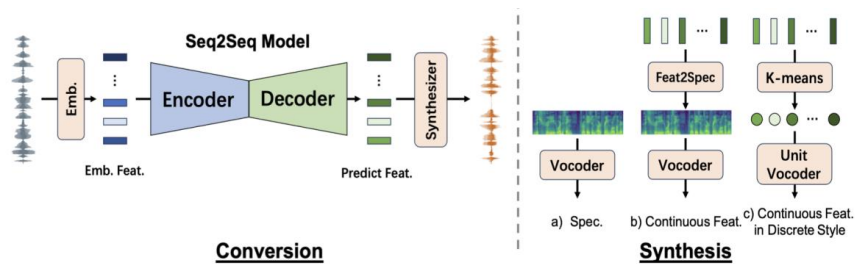


Fig. 4.4 Architecture of Sea-to-Sea VC

In Fig. 3, Seq-to-Seq VC (Fig. 4.4) [29] is a model used to convert one speaker's voice into another's by learning a mapping between the source and target feature representations. In this case, the source feature is PPG, and the target feature is Mel-spectrogram. PPG provides a phoneme-level representation of audio, capturing the phonetic information of speech. The Mel spectrogram, on the other hand, represents the frequency content of the speech signal using a non-linear frequency scale (Mel scale), containing both accent identity and speech content. The model uses an encoder-decoder architecture. The encoder receives the source PPG features and encodes them into a high-dimensional feature space. The latent representation generated by the encoder captures content and speaker-independent information. The decoder then takes the latent representation and generates the corresponding Mel-spectrogram frames.

The model is trained on paired data, where each PPG feature is paired with its corresponding Mel-spectrogram. During training, the model learns to map PPG features to Mel-spectrogram features by minimizing the difference between predicted and actual Mel-spectrograms. Once trained, the model can be used for voice conversion. For a new speech sample, the model receives the source PPG features and generates the target Mel-spectrogram, which is then converted into a waveform using a vocoder. This way, the model transforms the accent identity while preserving the speech content.

In the current study, this framework was adapted to convert native speaker audio into imitation audio produced in the learner's voice. Specifically, native speakers' audio was used as the source data, while the corresponding learner's audio was used as the target data. The seq-to-seq model was trained with these paired datasets, allowing it to learn how each learner would naturally reproduce the native speaker's speech. To further enhance the naturalness and accuracy of the synthesized speech, a separate vocoder was individually trained for each learner. In this case, the Parallel WaveGAN [30] model

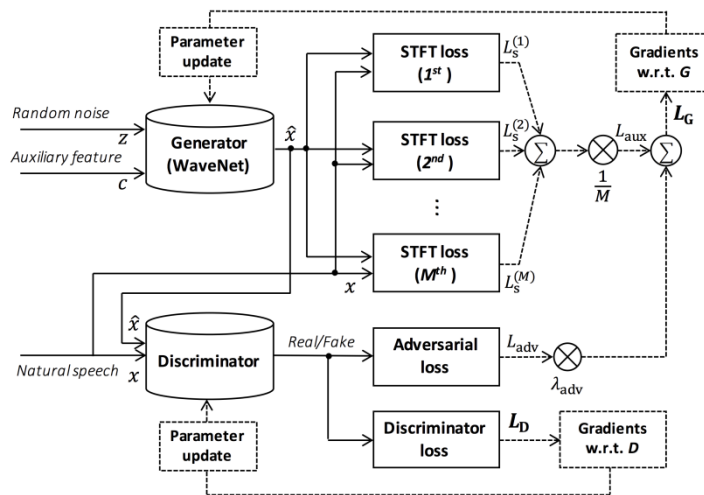


Fig. 4.5 Architecture of Parallel WaveGAN

was employed as the vocoder.

Parallel WaveGAN (Fig. 4.5) is a speech synthesis model based on Generative Adversarial Networks (GAN) [31], designed to convert acoustic features such as Mel-spectrograms into high-quality time-domain waveforms. The model consists of two main components: the generator and the discriminator. The generator is responsible for generating audio waveforms based on the input Mel-spectrogram, while the discriminator guides the generator by determining whether the generated audio is real or fake. The generator learns the mapping between Mel-spectrograms and time-domain waveforms through a convolutional neural network (CNN) [32], while the discriminator uses another CNN to distinguish between real and generated audio. During training, the generator and discriminator undergo adversarial training, where the generator aims to produce waveforms that are indistinguishable from real audio, and the discriminator works to differentiate real from fake audio. The generator is optimized by minimizing the difference between the generated waveform and the real waveform. Parallel WaveGAN's parallel generation approach allows for fast inference while maintaining high audio quality, making it suitable for applications in real-time speech synthesis.

This personalized approach enabled the production of imitation speech tailored to each learner, supporting a more precise analysis of prosody imitation and accent adaptation.

### 4.2.2 Data Processing

Since training a learner's model takes about a week, it is not feasible to train a model for each learner. Therefore, we selected a few learners with different performances as representatives for training. We calculated the mean phonetic difference between each learner and the model audios, and sorted the learners based on their phonetic imitation performance. Learners were then divided into three equal groups based on their performance: the group with the lowest differences was classified as *Good*, the middle group as *Medium*, and the other group as *Poor*. Subsequently, two learners were randomly selected from each group, and their audio were used for analysis. Here are the selected learners for each group:

Good performance: (G1, G2)

Medium performance: (M1, M2)

Poor performance: (P1, P2)

According to previous research [33], learners gradually improved their imitation ability through practice in the STEAC course. To simulate this trend using the model, the Seq-to-Seq model was trained with datasets divided in a way that reflects the learners' progress over time. Specifically, the datasets were organized into different stages, capturing the evolution of learners' imitation skills as they continued their practice. This approach allowed the model to mimic the learners' gradual improvement,

providing a more realistic simulation of their development and refining the prosody imitation process over time. The Seq-to-Seq model was trained with datasets divided as follows:

- Dataset I: Train the model using data from Week 2 to Week 4 to predict the results for Week 5.
- Dataset II: Train the model using data from Week 3 to Week 5 to predict the results for Week 6.
- Dataset III: Train the model using data from Week 4 to Week 6 to predict the results for Week 7.

Table 4.1 Division of the Training Dataset

	Dataset I	Dataset II	Dataset III
Week2 (0.8x)	Training	—	—
Week3 (0.9x)	Training	Training	—
Week4 (1.0x)	Training	Training	Training
Week5 (1.0x)	Predicting	Training	Training
Week6 (1.0x)	—	Predicting	Training
Week7 (1.0x)	—	—	Predicting

Additionally, a separate vocoder was trained for each of the six learners using all of their audio data. For this, the Parallel WaveGAN model was used. Starting from the Week2 to Week7, all the audio data of the selected learners were used as input to train the vocoder.

Since the VC model requires at least three weeks of learner data for training, it can be used for difficulty prediction after the course has been ongoing for about three weeks.

## 4.3 TTS

### 4.3.1 Introduction of ElevenLabs TTS

Text-to-Speech (TTS) is a technology that converts written text into spoken audio [34]. Here we used ElevenLabs TTS for analysis. ElevenLabs TTS [35] is an advanced text-to-speech system that uses deep learning models to generate natural and expressive speech. It effectively captures context, tone, and emotion, making the output sound more human-like. With powerful voice cloning technology, it can replicate voices using minimal data, offering high flexibility for personalization. Its end-to-end processing and high-quality neural vocoder ensure smooth and realistic audio generation, making it suitable for various applications like audiobooks, virtual assistants, and content creation.

The ElevenLabs TTS model is an advanced speech synthesis model known for producing highly natural and expressive audio. It is allowed for users to upload their audio recordings to implement their own TTS models. There are two versions of the ElevenLabs TTS: **Professional** and **Instant**.

The Professional mode of ElevenLabs can generate higher-quality audio compared to the Instant mode. It allows uploading up to 100 samples, with each sample not exceeding 10MB. Additionally, for identity verification, users must read aloud a displayed sentence within a 15-second time limit. This mode requires approximately one hour of fine-tuning to train the model, resulting in more accurate and natural-sounding speech synthesis.

In contrast, the Instant mode produces audio of moderate quality. It permits uploading up to 25 samples, each also limited to 10MB, with a total duration of at least 5 minutes. This mode does not require user verification, and model training does not involve fine-tuning. Instead, it generates a target speaker's TTS model in about 2 minutes, offering faster but less refined results.

Since the STEAC course utilized lecture recordings for analysis, obtaining learners' personal verification was challenging. As a result, the Professional mode could not be implemented in our experiments. Therefore, in this research, we selected the Instant mode for training the TTS models for individual learners, enabling efficient model generation without the need for direct user verification.

### 4.3.2 Data Processing

When utilizing the ElevenLabs TTS model, we employed the same datasets that were previously used for the VC model to maintain consistency in data comparison and analysis (Dataset I, Dataset II and Dataset III).

Specifically, for each of the six selected learners, we uploaded their corresponding weekly audio data (around 18 minutes) into the ElevenLabs platform using the *Instant*

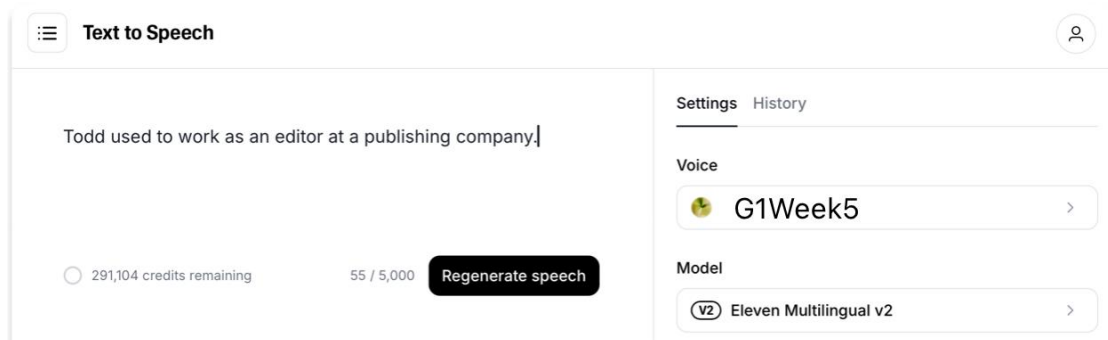


Fig. 4.6 An example for generating audio use ElevenLabs TTS

mode. In this mode, we specified the accent : *Japanese* to tailor the model's speech output to the desired linguistic characteristics.



After training the model, select the Text to Speech mode, input the text to be converted into audio, and choose the trained voice model to generate the corresponding audio output (Fig. 4.6).

By selecting the model corresponding to each learner's dataset and sequentially inputting the texts of all audio files used in week 5, week 6 and week 7, all TTS-predicted imitation results can be obtained.

Similar to the VC model, the TTS model also requires at least three weeks of learner audio data to ensure the accuracy of its predictions. Therefore, the TTS model is also suitable for difficulty prediction after the first three weeks of the course.

### 4.4 Experiment

The GSLM model only takes native speaker audio as input and does not require any learner audio. In contrast, the training of TTS models requires individual learners' audio, and any text can be converted to speech with the learners' voice quality. The VC model requires audio of both a learners and natives for training, and it maps native audio to its corresponding speech with the learners' voice quality and style.

Due to differences in input content and format, it can be predicted that the VC and TTS models can provide individualized predictions for each learner, whereas the GSLM model can only offer a general prediction of overall difficulty for learners simply because learners' data are not used at all. Further, since the TTS model relies solely on text input and lacks the prosodic features of native speaker audio, its prediction performance may be less effective in certain tasks, compared to the VC models.

In the experiments, predictions were made using the three models: GSLM, TTS, and VC. There are two main experimental objectives in this section:

- (a) determining whether those three models can effectively function as a learner simulator,
- (b) evaluating whether it can be utilized to estimate the degree of difficulty learners may experience when imitating new speech material.

To assess the models' performance in simulating individual learners (objective a), the correlations between predicted prosody and actual prosody observed in the learners' recordings were calculated. This process involves evaluating how closely the prosodic features (such as pitch, duration, and intensity) of the predicted audio match the prosody in the actual recordings made by the learners. Here, STFT and SPTK were utilized to extract intensity and pitch from all audio samples, also DTW and FA were used to align the audio samples.

To achieve evaluating whether it can be utilized to estimate the degree of difficulty learners may experience when imitating new speech material (objective b), this chapter primarily uses scatter plots of XX (XX = corr. (Learner, Native)) and YY (YY = corr.

(generated Audio, Native)), and calculates the correlation coefficients (similarity).

#### 4.4.1 GSLM models and their performance

To achieve objective (a), correlation coefficients for vowel prosody control—specifically focusing on pitch, duration, and intensity—were calculated between the learners' actual speech recordings and the audio produced by the GSLM model. These correlation values, presented in Table 4.2, provide quantitative insights into how closely the GSLM's generated output aligns with the prosodic patterns of learner speech.

Table 4.2 Correlation coefficients between actual learners and generated audio (GSLM)

	Pitch	Intensity	Duration
JP1000	0.50	0.41	0.82
JP200	0.53	0.37	0.76
JP50	0.38	0.31	0.78
mJP1000	0.51	0.48	0.69
mJP200	0.52	0.44	0.67
mJP50	0.49	0.40	0.47

The analysis revealed that the GSLM model demonstrates a reasonable degree of similarity to learners' real speech, particularly in the aspects of pitch and duration, which are critical components of natural and fluent speech production. This suggests that the model is capable of capturing and replicating some of the prosodic features that learners naturally exhibit. Such findings indicate the GSLM's potential as a useful tool for simulating learner speech behavior and for predicting which speech materials might pose greater challenges for learners to imitate, thereby aiding in the design of more effective pronunciation training and language learning programs.

For the modified GSLM, a component was added to alter the generated audio length. However, its generated results showed lower similarity in duration compared to the learner's audio than the baseline model, which did not introduce duration variation. For overlapping courses, students will strive to mimic the model speech audio, which results in some prosody deviations in their utterances. As a consequence, the modified GSLM results, which generated natural Japanese accent English, may be lower than the baseline GSLM results.

Next, to explore whether GSLM can estimate the difficulty of prosody imitation, which may vary among learners, an overall analysis is conducted for each approximately 30-second passage. The analysis is based on comparison between the

passage-based means of two kinds of correlations, one is between actual learners' prosody and native prosody, and the other is between predicted prosody and native prosody. If the former mean is lower, the passage is difficult to imitate, and vice versa. If the similarity between the two kinds of correlations is higher, we can claim the model can estimate the difficulty of prosody imitation in new material well. Figs 4.7, 4.8, 4.9, and 4.10 are results of verifying how well the GSLMs can estimate the difficulty of imitation.

### 4.4.1.1 Difficulty Estimation for Pitch Control

To estimate the degree of difficulty learners may experience when imitating new speech material, scatter plots were created, and correlation coefficients were calculated for analysis.

We plotted scatter plots of the pitch correlation coefficients between the generated audio and the native speakers' audio, as well as between the learners' audio and the native speakers' audio. Additionally, we calculated the correlation between the generated audio and the learners' audio in terms of pitch control. The x-axis represents the correlation coefficients between the learners' audio and the native speakers' audio, while the y-axis represents the correlation coefficients between the audio generated by JP200/mJP200 and the native speakers' audio, as well as JP1000/mJP1000 and the native speakers' audio.

From Figs. 4.7 and 4.8, it can be observed that the four types of GSLM models effectively estimate the difficulty learners may face when imitating the pitch patterns

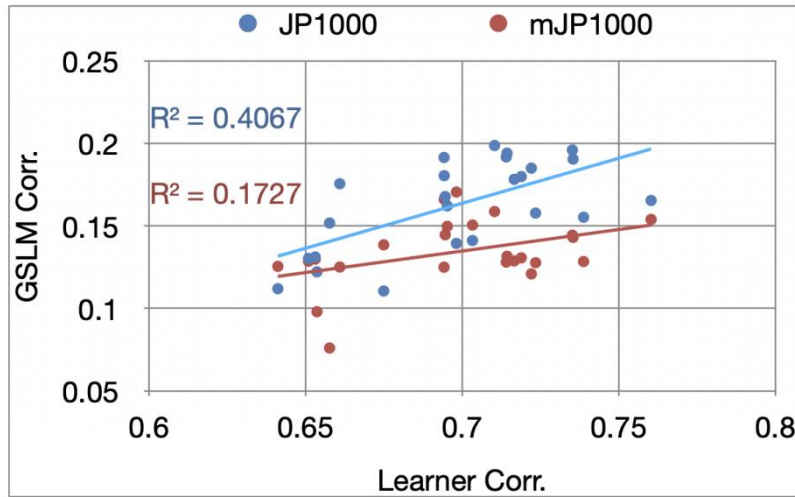


Fig. 4.8 Pitch Correlation Coefficients ((m)JP1000 Learners)

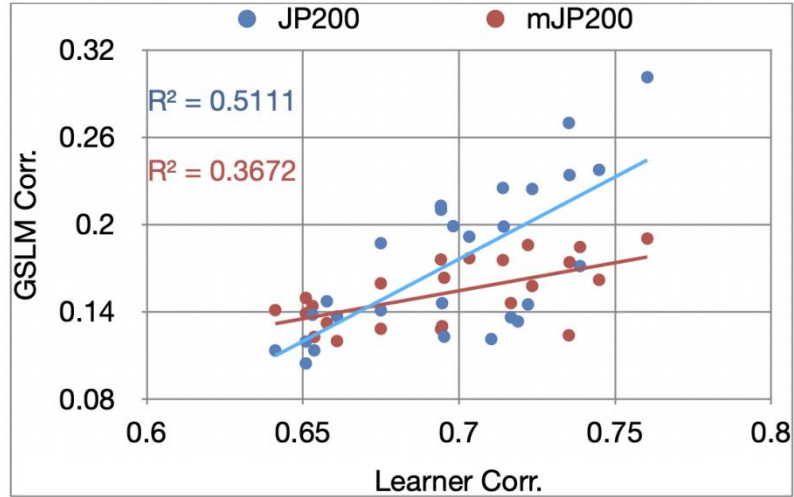


Fig. 4.7 Pitch Correlation Coefficients ((m)JP200 Learners)

of new materials, with similarities ranging from 0.42 to 0.71. The highest similarity of 0.71 for JP200 indicates better pitch imitation performance.

#### 4.4.1.2 Difficulty Estimation for Intensity Control

We also plotted scatter plots of the intensity correlation coefficients between the generated audio (JP200, mJP200/ JP1000, mJP1000) and the native speakers' audio, as well as between the learners' audio and the native speakers' audio. Additionally, we calculated the correlation between the generated audio (JP200, mJP200/ JP1000, mJP1000) and the learners' audio in terms of intensity control. The x-axis represents the correlation coefficients between the learners' audio and the native speakers' audio, while the y-axis represents the correlation coefficients between the audio generated by JP200, mJP200 and the native speakers' audio, as well as JP1000/mJP1000 and the

native speakers' audio.

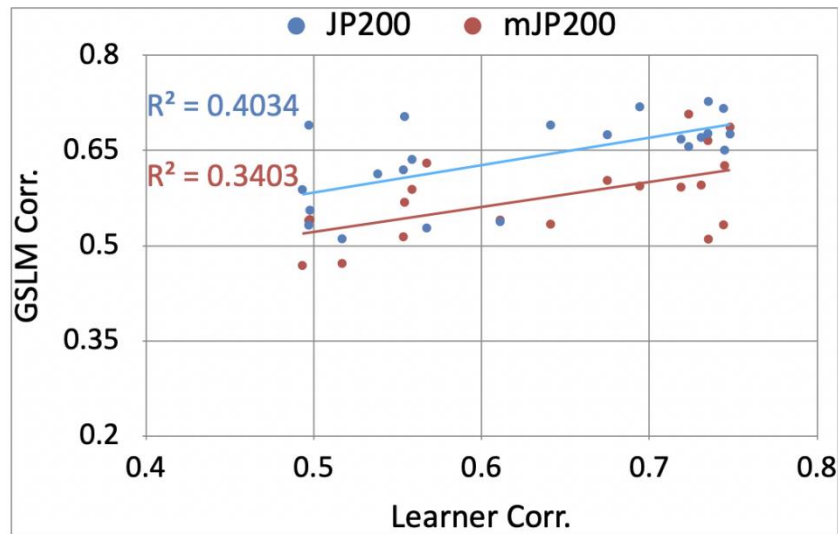


Fig. 4.9 Intensity Correlation Coefficients ((m)JP200 Learners)

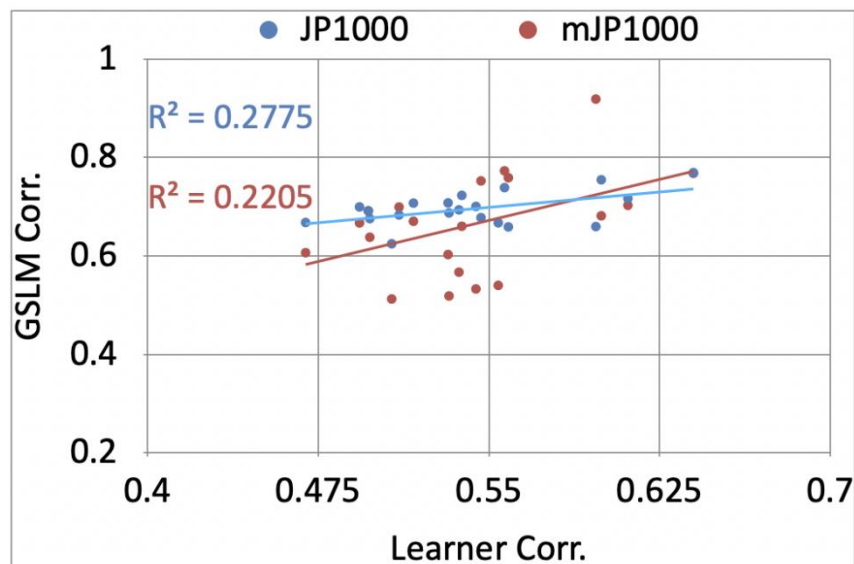


Fig. 4.10 Intensity Correlation Coefficients ((m)JP1000 Learners)

For the intensity results shown in Figs. 4.9 and 4.10, all four models continued to perform well in estimating imitation difficulty, with similarities ranging from 0.47 to 0.63. Once again, JP200 exhibited the highest similarity.

#### 4.4.1.3 Difficulty Estimation for Duration Control

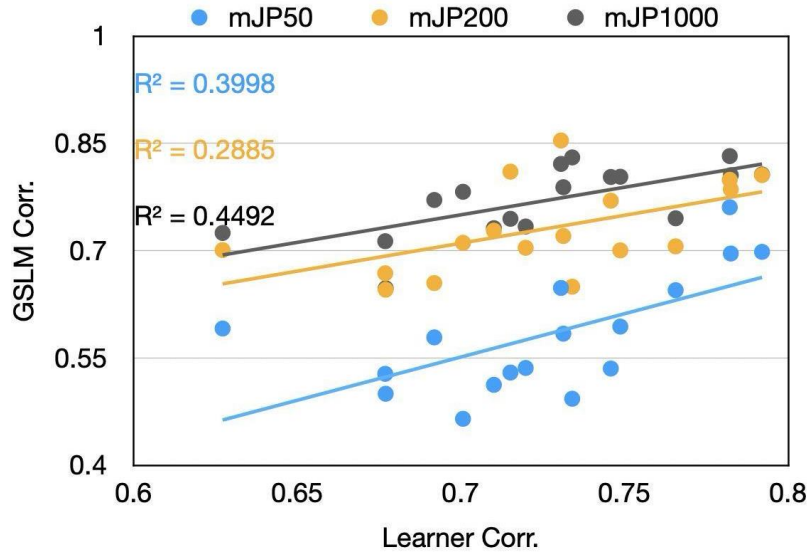


Fig. 4.11 Duration-based difficulty estimation in the modified GSLM

Since the original GSLM is a model that directly maps frame-by-frame, the duration difference between the generated audio and the input audio is minimal, so difficulty estimation for duration control is analyzed only using the modified GSLM.

Fig. 4.11 shows the results, from which, it can be observed that the mJP50, mJP200, and mJP1000 models all exhibit good performances of estimating the difficulty level of duration-based prosody imitation. The highest similarity is 0.67 found in the case of mJP1000.

#### 4.4.2 VC/TTS models and their performance

Table 4.3. Correlation coefficients between actual learners and generated audio (VC & TTS)

	Pitch	Intensity	Duration
VC model	0.52	0.60	0.78
TTS model	0.49	0.62	0.75

As with the previous case, the correlation coefficients between the audio generated by the VC and TTS models and the actual learner's audio were calculated here to measure the similarity between the generated audio and the learner's audio.

Table 4.3 presents the results, and these results indicate that the VC models slightly outperform the TTS models in pitch and duration control, suggesting that the VC models are more effective in capturing and replicating the learners' prosodic patterns in these aspects. This advantage is likely due to the VC models' ability to directly utilize

both the learners' and native speakers' audio during training, allowing for more accurate modeling of temporal and pitch variations.

Conversely, the TTS models demonstrate slightly better precision in intensity control. This may be because TTS models, which are designed to synthesize speech from text, can produce smoother and more consistent intensity patterns, despite lacking direct access to detailed acoustic features during training.

When comparing the VC/TTS models with the GSLM models presented in Table 4.2, the VC and TTS models show significantly higher performance in simulating learners' prosody imitation, particularly in intensity control. This substantial improvement suggests that both the VC and TTS models are more effective than the GSLM models in capturing the dynamic intensity variations that learners struggle to imitate.

Next, the performance of estimating the difficulty level of prosody imitation is examined separately for duration, pitch, and intensity control. Scatter plots and correlation coefficient calculations were again used to evaluate the accuracy of difficulty prediction. Figs (4.12 ~ 4.20) show scatter plots of the data from one learner (G1). The x-axis represents the correlation coefficients between the learners' audio and the native speakers' audio, while the y-axis represents the correlation coefficients between the audio generated by the VC or TTS model and the native speakers' audio.

Figs (4.12 ~ 4.14) illustrate the prediction results for pitch, intensity, and duration in Week 5 (Training data: Week2 ~Week4, Predicting: Week5). Figs (4.15 ~ 4.17) show the results for Week 6(Training data: Week3 ~Week5, Predicting: Week6), and Figs (4.18 ~ 4.20) present the results for Week 7(Training data: Week4 ~Week6, Predicting: Week7).

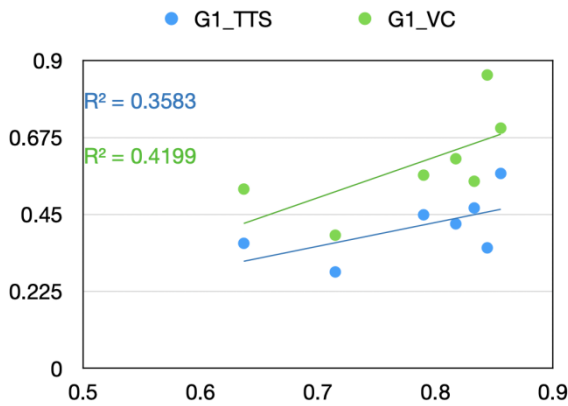


Fig. 4.12 Pitch Correlation Coefficients for Week5

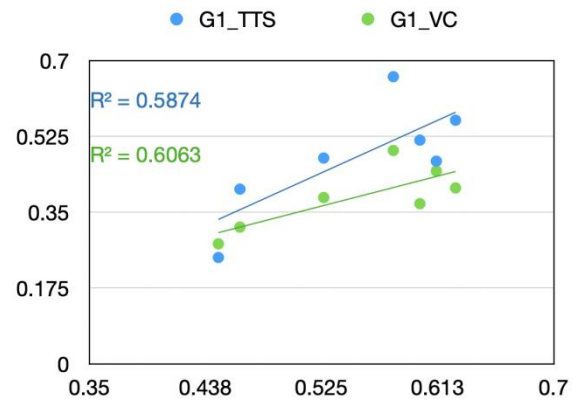


Fig. 4.13 Intensity Correlation Coefficients for Week5

## 4. Modeling and Prediction

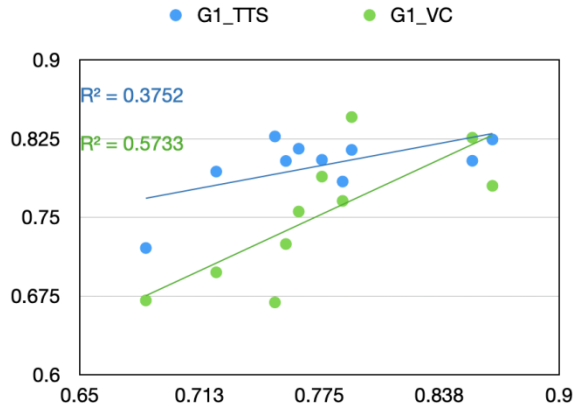


Fig. 4.14 Duration Correlation Coefficients for Week5

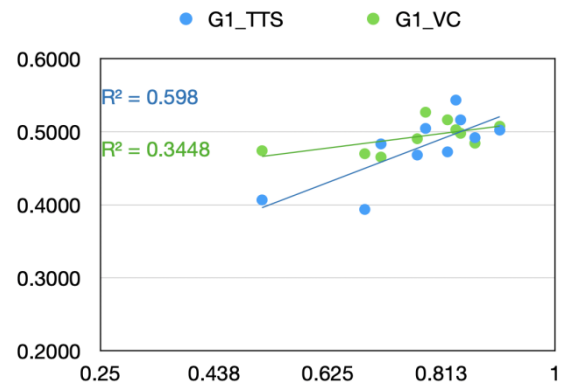


Fig. 4.15 Pitch Correlation Coefficients for Week6

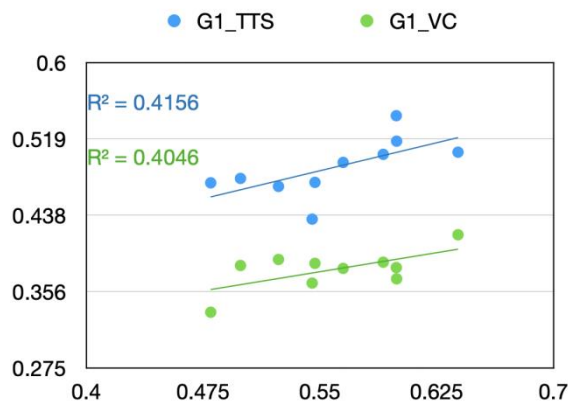


Fig. 4.16 Intensity Correlation Coefficients for Week6

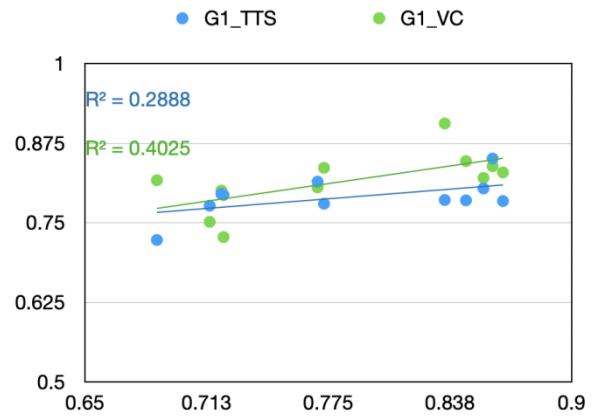


Fig. 4.17 Duration Correlation Coefficients for Week6

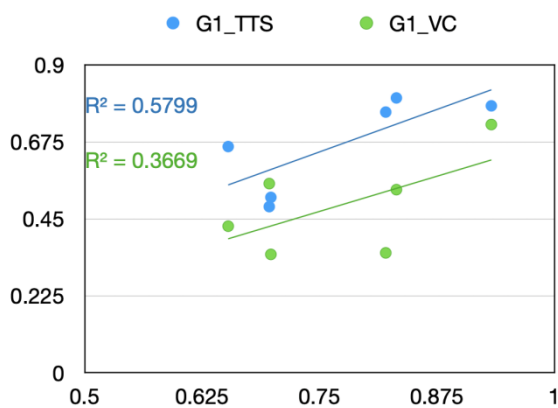


Fig. 4.18 Pitch Correlation Coefficients for Week7

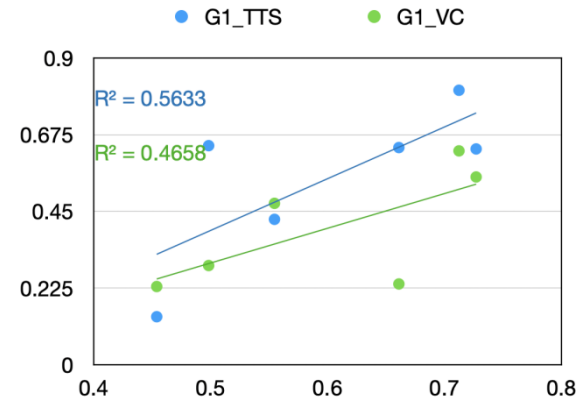


Fig. 4.19 Intensity Correlation Coefficients for Week7



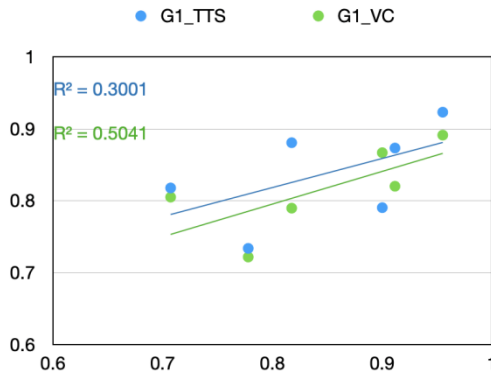


Fig. 4.20 Duration Correlation Coefficients for Week7

Subsequently, the calculated results were averaged within each learner group—specifically, the results of (G1, G2), (M1, M2), and (P1, P2) were averaged. The summarized results are presented in Tables 4.4, 4.5, and 4.6. The detailed original data tables for each learner are provided in the Appendix A.

#### 4.4.2.1 Difficulty Estimation for Duration Control

In terms of duration, as shown in Table 4.4, the VC model demonstrated superior prediction accuracy compared to the TTS model. This difference in performance can be attributed to the distinct training methods of the two models. The VC model is trained using both the native speaker's audio and the learner's audio, allowing it to directly capture and model the prosodic features, including subtle variations in duration, more effectively. By leveraging actual speech data from both sources, the VC model

Table. 4.4 Duration-based difficulty estimation

	Week5	Week6	Week7
G_TTS	0.55	0.58	0.61
G_VC	0.73	0.63	0.68
M_TTS	0.48	0.45	0.56
M_VC	0.58	0.51	0.53
P_TTS	0.48	0.50	0.81
P_VC	0.59	0.61	0.78

can generate outputs that more closely mimic the native speaker's speech patterns, leading to higher accuracy in predicting the difficulty level of prosody imitation. On the other hand, the TTS model relies only on text-based input from both the learner and the native speaker, which limits its ability to fully capture the nuances of speech duration. Since text input lacks detailed prosodic information, the TTS model struggles to accurately reproduce the temporal aspects of speech, resulting in lower prediction performance.

### 4.4.2.2 Difficulty Estimation for Pitch and Intensity Control

When examining pitch and intensity in Tables 4.5 and 4.6, the results show considerable variation depending on the learners' performance levels and the training data used. In Week 5, the VC model outperformed the TTS model in both pitch and intensity prediction. This advantage may stem from the VC model being trained on data from Week 2 (0.8×), Week 3 (0.9×), and Week 4 (1.0×), whereas the prediction data for Week 5 is at the standard rate of 1.0×. This inconsistency in training and prediction data

Table 4.5 Pitch-based difficulty estimation

	Week5	Week6	Week7
G_TTS	0.63	0.67	0.66
G_VC	0.75	0.55	0.52
M_TTS	0.41	0.51	0.50
M_VC	0.49	0.48	0.61
P_TTS	0.42	0.65	0.38
P_VC	0.57	0.91	0.49

Table 4.6 Intensity-based difficulty estimation

	Week5	Week6	Week7
G_TTS	0.71	0.64	0.78
G_VC	0.74	0.55	0.73
M_TTS	0.60	0.42	0.65
M_VC	0.71	0.61	0.65
P_TTS	0.75	0.55	0.61
P_VC	0.96	0.75	0.79

likely impacted the models' performance, particularly affecting the TTS model, which does not incorporate prosodic information from native speaker audio in its input. Without direct access to acoustic features, the TTS model may struggle to generalize effectively to new speech data, leading to lower prediction accuracy.

In contrast, during Weeks 6 and 7, a distinct performance pattern emerged based on the learners' proficiency. For learners with Good performance, the TTS model achieved higher prediction accuracy than the VC model in both pitch and intensity control. This suggests that the TTS model may be better suited for learners who already exhibit relatively stable prosodic patterns, possibly due to its text-to-speech synthesis approach, which produces smoother and more consistent prosodic features.

Conversely, for learners with Poor performance, the VC model demonstrated superior prediction accuracy compared to the TTS model. One possible explanation for this outcome lies in the characteristics of the instant mode in the Eleven Labs TTS tool. The instant mode is designed to operate with only a small number of training samples, suggesting that it may rely on pre-registered native speech samples during synthesis. As a result, the speech generated by the instant mode tends to have lower accentedness than expected. This unintended reduction in accentedness could make the TTS model less effective at capturing the prosodic deviations typically found in learners with lower performance.

In contrast, the VC model, trained directly on both the learners' and native speakers' audio, is more capable of reflecting the variability and imperfections present in the learners' speech. This allows it to provide more accurate predictions for learners who struggle with prosody imitation.

## 4.5 Results and Discussion

For the GSLM model, it has been confirmed that it can generate results similar to the learners' data in terms of pitch control and intensity control. And also for the modified GSLM model, it showed high similarity with learner data in duration control. The GSLM model is simple to train and does not require input from the learner's audio, making it suitable for predicting the general difficulty level for all learners. Specifically, the baseline GSLM is used to predict the difficulty of mimicking pitch and intensity, while the modified GSLM is applied to predict the difficulty of controlling duration.

The VC and TTS models, on the other hand, are suitable for predicting individual learners. With approximately three weeks of learner audio data available, the VC model can be used to predict duration or to predict pitch and intensity control for tasks involving different speech speeds. For learners with better oral proficiency, the TTS model can be used to predict tasks where the speech rate remains largely unchanged. However, for learners with medium or lower oral proficiency, the VC model is recommended for predicting pitch and intensity control.

In general, before the course begins, the baseline GSLM (JP200) model can be used to predict the difficulty of pitch and intensity, while the modified GSLM (mJP1000) model can be used to predict the difficulty of duration control. Simpler speech materials can be selected for the first three weeks of the course. After the three weeks, learners' performance can be sorted. If there is little difference in speech rate between the training and prediction data, the TTS model can be used for predicting pitch and intensity difficulty for learners with good performance, and the VC model can be used for predicting duration control difficulty. For learners with medium or poor performance, the VC model can be used to predict pitch, intensity, and duration difficulty.

In cases where there is a large difference between the training and prediction data (e.g., STEAC 2023S Week 5), the VC model should be used to predict the pitch, intensity, and duration control difficulty for all learners. After obtaining the prediction results, slightly more challenging practice materials can be selected for the learners in the following week.

# Chapter 5

## Conclusions and Future Works

### 5.1 Conclusions

In conclusion, this paper, through a prosodic analysis of the overlapping audio from the STEAC 2023S learners, has revealed several key insights. It was found that learners struggle particularly with mimicking the duration of longer vowels and with achieving optimal results in sentences that start with high intensity. Furthermore, learners tend to perform less well in tasks that involve rapid pitch changes.

The study also examined the effectiveness of different models for predicting prosodic features. The GSLM model was found to accurately replicate the learners' pitch control and intensity control. In its modified version, the GSLM model also showed a high degree of similarity with learner data in terms of duration control. Due to its simplicity and the fact that it does not require learner-specific audio input, the GSLM model is deemed effective for predicting the general difficulty level across all learners.

On the other hand, the VC and TTS models are more suited for predicting individual learner performance. With approximately three weeks of learner audio data available, the VC model was found to be useful for predicting duration or for estimating pitch and intensity control in tasks involving different speech speeds. For learners with higher oral proficiency, the TTS model was more effective in predicting tasks where the speech rate remains relatively constant. However, for learners with medium or lower oral proficiency, the VC model was recommended for more accurate predictions of pitch and intensity control.

Ultimately, these findings highlight the potential of using GSLM, VC, and TTS models to support tailored learning materials and to predict learner performance in English oral proficiency tasks, offering valuable insights for enhancing learner-specific training.

### 5.2 Future Work

To improve the accuracy and efficiency of the prediction models, several areas can be targeted for optimization:

1. Enhancing Prediction Accuracy: The current models still have limitations when it

## 5. Conclusions and Future Works

---

comes to mimicking the learner's speech, and the prediction accuracy is not as high as desired. Future work will focus on refining these models to better capture the nuances of individual learners' prosody, aiming to achieve more precise predictions in pitch, intensity, and duration control.

2. Simplifying the Training Process: The training process of current models, particularly the seq-to-seq VC model, is time-consuming and requires individual training for each learner. This involves training both the seq-to-seq model and the vocoder separately, with each learner's model taking approximately a week to train. To overcome this, efforts will be made to streamline the training workflow and reduce the time required to train models, making it more feasible to train separate models for a larger number of learners.

3. Expanding Learner Diversity: This study currently focuses solely on the English speech of Japanese learners. Future research will involve collecting and analyzing data from learners of various nationalities to identify the specific pronunciation challenges faced by speakers of different native languages and to select the most suitable prediction models for accurate speech prediction.

By optimizing these aspects, the models can become more efficient and applicable for large-scale use, enabling personalized learning experiences for a broader range of learners.

## Reference

- [1] Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
- [2] Hisagi, M., Nishi, K., & Strange, W. (2008). Acoustic properties of Japanese and English vowels: Effects of phonetic and prosodic context. *Japanese/Korean Linguistics*, 13, 223-224.
- [3] Roach, Peter (2004), "British English: Received Pronunciation", *Journal of the International Phonetic Association*, 34 (2): 239–245, doi:10.1017/S0025100304001768 (inactive 2024-11-01)
- [4] Okada, Hideo (1999), "Japanese", in International Phonetic Association (ed.), *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, pp. 117–119, ISBN 978-0-52163751-0
- [5] Arciaga, Kasumi & Flores, Eden. (2021). Phonological features of Japanese speakers of English: A descriptive analysis.
- [6] Makino, T. (2014). Pronunciation characteristics of Japanese speakers' English: A preliminary corpus--based study. In J. Levis & S. McCrocklin (Eds). *Proceedings of the 5th Pronunciation in Second Language Learning and Teaching Conference* (pp.121--136). Ames, IA: Iowa State University.
- [7] Ohata, K. (2004). Phonological Differences between Japanese and English: Several Potentially Problematic Areas of Pronunciation for Japanese ESL/EFL Learners. *Asian EFL Journal* 6, 1-19.
- [8] Saito, K. (2011). Identifying problematic segmental features to acquire comprehensible pronunciation in EFL settings: The case of Japanese learners of English. *RELC Journal* 42(3). 363-378.
- [9] MacCathy, P. (1978). *The teaching of pronunciation*. Cambridge: Cambridge UP.
- [10] Clark John, Yallop Collin, Fletcher Janet (2007). *Introduction to Phonetics and Phonology*. Oxford: Blackwell. pp. (pp)340.
- [11] Y. Shibuya, "Differences between native and non-native speakers' realization of stress-related durational patterns in American English," in *Journal of the Acoustical Society of America*, vol. 100, 1996, p. 2725.
- [12] T. Nariai and K. Tanaka, "A study on pitch patterns in Japanese speakers of English with verification by speech re-synthesis," in *IEICE Transactions on Information and Systems*, vol. E94.D, no. 12, pp. 2495–2502, 2011.
- [13] C. Shoda, Y. Gao, Y. He, N. Minematsu, D. Saito, and N. Nakanishi, "Learners' Prosodic Control in the Task of Expressive Storytelling and Predicted Native Listeners' Impressions of the Learners' Speech," in *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023.

- 
- [14] N. Minematsu, N. Nakanishi, Y. Gao, and J. Choi, "Karaoke shadowing training to facilitate language learners to acquire L2 prosodic control," in Proc. ICPHS, 2023.
  - [15] E. J. Keogh and M. Pazzani, "Derivative Dynamic Time Warping," in SDM 2001, Society for Industrial and Applied Mathematics.
  - [16] Prasanna, S. & Gangashetty, Suryakanth & Yegnanarayana, B. (2002). Significance Of Vowel Onset Point For Speech Analysis.
  - [17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motl'ı'cek, Y. Qian, P. Schwarz, J. Silovsk'y, G. Stemmer, and K. Vesel'y, "The Kaldi speech recognition toolkit," in Proc. Automatic Speech Recognition and Understanding, 2011.
  - [18] Jeon, Hohyub & Jung, Yongchul & Lee, Seongjoo & Jung, Yunho. (2020). Area-Efficient Short-Time Fourier Transform Processor for Time-Frequency Analysis of Non-Stationary Signals. Applied Sciences. 10. 7208. 10.3390/app10207208.
  - [19] Speech Signal Processing Toolkit (SPTK). [Online]. Available: <https://sp-tk.sourceforge.net>
  - [20] T. Riney and J. Anderson-Hsieh, "Japanese Pronunciation of English," JALT Journal, Vol. IS, No.1 (May 1993)
  - [21] K. Arciaga and E. Flores, "Phonological features of Japanese speakers of English: A descriptive analysis," 2021.47
  - [22] K. Lakhotia et al., "On generative spoken language modeling from raw audio," Transactions of the Association for Computational Linguistics, vol. 9, pp. 1336–1354, 2021.
  - [23] A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.
  - [24] W.-N. Hsu et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.
  - [25] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel-spectrogram predictions," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783, 2018.
  - [26] Onda, K., Park, J., Minematsu, N., & Saito, D. (2024). "A Pilot Study of GSLM-based Simulation of Foreign Accentuation Only Using Native Speech Corpora". *arXiv preprint arXiv:2407.11370*.
  - [27] 恩田健太郎, 朴浚鎔, 井本桂右, 深山覚, 齋藤大輔, 峯松信明, "離散トークンの継続長予測に基づく母語話者音声コーパスのみを用いた外国語訛り音声合成手法の改善," 研究報告音声言語情報処理 (SLP), vol. 2024-SLP-154, no. 22, pp. 1-7, Dec. 5, 2024.



- 
- [28] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.-Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, International Conference on Learning Representations, (online), available from <https://openreview.net/forum?id=piLPYqxtWuA> (2021).
  - [29] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," IEEE/ACM TASLP, vol. 29, pp. 745–755, 2021.
  - [30] Yamamoto, R., Song, E., & Kim, J. (2019). Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6199-6203.
  - [31] M. Krichen, "Generative Adversarial Networks," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-7, doi: 10.1109/ICCCNT56998.2023.10306417.
  - [32] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.
  - [33] Y. Gao, J. Choi, N. Minematsu, N. Nakanishi, D. Saito, "Automatic Prediction of Language Learners' Listenability Using Speech and Text Features Extracted from Listening Drills". Proc. Interspeech 2023, 979-983, doi: 10.21437/Interspeech.2023-2541.
  - [34] P. Taylor, Text-to-Speech Synthesis. Cambridge, U.K.: Cambridge University Press, 2009.
  - [35] ElevenLabs, "ElevenLabs – The Prime Voice AI," ElevenLabs. [Online]. Available: <https://elevenlabs.io>.
  - [36] Fromkin, V., Rodman, R., & Hyams, N.M. (2003). *An Introduction to Language* (p. 236). Thomson/Heinle. ISBN: 9780155084810.

# Appendix A

## Tables

Table 1. Pitch-based pitch estimation

Pitch	Week5	Week6	Week7
G1_TTS	0.60	0.77	0.76
G1_VC	0.65	0.59	0.61
G2_TTS	0.65	0.57	0.56
G2_VC	0.85	0.52	0.43
M1_TTS	0.44	0.50	0.54
M1_VC	0.53	0.49	0.66
M2_TTS	0.37	0.52	0.45
M2_VC	0.45	0.46	0.56
P1_TTS	0.46	0.78	0.38
P1_VC	0.54	0.96	0.42
P2_TTS	0.39	0.52	0.38
P2_VC	0.61	0.86	0.56

Table 2. Intensity-based intensity estimation

Intensity	Week5	Week6	Week7
G1_TTS	0.77	0.66	0.75
G1_VC	0.78	0.64	0.68
G2_TTS	0.65	0.62	0.80
G2_VC	0.70	0.46	0.77
M1_TTS	0.72	0.49	0.80
M1_VC	0.74	0.79	0.78
M2_TTS	0.49	0.35	0.50
M2_VC	0.68	0.44	0.51
P1_TTS	0.86	0.58	0.49
P1_VC	0.98	0.63	0.76
P2_TTS	0.64	0.52	0.74
P2_VC	0.94	0.86	0.82

Table 3. Duration-based duration estimation

Duration	Week5	Week6	Week7
G1_TTS	0.61	0.54	0.55
G1_VC	0.76	0.63	0.71
G2_TTS	0.49	0.62	0.67
G2_VC	0.71	0.63	0.66
M1_TTS	0.46	0.48	0.50
M1_VC	0.61	0.50	0.64
M2_TTS	0.51	0.42	0.62
M2_VC	0.55	0.52	0.42
P1_TTS	0.50	0.61	0.71
P1_VC	0.52	0.74	0.78
P2_TTS	0.47	0.39	0.91
P2_VC	0.66	0.48	0.77

## **Appendix B**

### **Publications**

#### **National conferences and meetings**

- Xiai Cheng, Kentaro Onda, Daisuke Saito and Nobuaki Minematsu, “Predicting Individual Language Learners' Oral Imitation By Modeling their Performance”. In 2024 Autumn Meeting of the Acoustical Society of Japan.
- Xiai Cheng, Haopeng Geng, Kentaro Onda, Daisuke Saito and Nobuaki Minematsu, “Modeling and Predicting Individual Learners' Performance of Prosody Imitation and its Use for Material Selection”. In 2025 Spring Meeting of the Acoustical Society of Japan.