E-015

A proposal for a voice conversion method using MelGAN-VC and TTS voices for listening practice of second-language learners

Kiyotada Mori † Yasuo Miyoshi † Ryo Okamoto †

Abstract

We propose a voice converter using MelGAN-VC which uses TTS voices generated by Amazon Polly as a training dataset. The source and target data consist of American English TTS voices and Japanese-accented ones defined as phoneme replacements, respectively. It is because this paper describes two preliminary experiments to focus on listening practice for Japanese ESL learners. The first experiment is to confirm Amazon Polly's Japanese-accented English speech is clear for Japanese learners. The second is to confirm MelGAN-VC which uses these data can convert American speech to Japanese-accented speech.

1. Introduction

Learning English involves 4 core skills: reading, writing, speaking, and listening. Among these skills, listening skills are one of the most important abilities because listening scores strongly affect other scores in English tests[1], i.e. strongly related to other skills. Nowadays, there are many ways to learn the language such as watching videos on video distribution services such as YouTube. In this situation, people who want to learn a new language often use speed adjustments to get used to listening to the language. An example of this learning method is to increase gradually playback speed, say from half to the original speed.

Using a method like that is called fadable scaffolding[2] in the field of cognitive science. The method is to reduce functions such as speed in the case of listening temporarily and revert functions gradually. Speed adjustment is a common method for learning a language. However, another potential method is an accent conversion for learning a language without changing audio speed.

This method is the so-called accent conversion, which is mostly used to convert a non-native speaker's accent to a native accent, making it more comprehensible for native speakers. However, we assert changing from the native speaker's accents to a listener's mother tongue accents is also effective for listening practice because its voice is closer to the listener's own voice.

There are various types of accent conversion[3], such as converting recognized voice using phoneme recognition or speech recognition into accented speech, which can be achieved through Text-To-Speech (TTS) systems. Additionally, another approach involves the use of deep neural networks, such as generative adversarial networks (GANs) or acoustic models, for accent conversion.

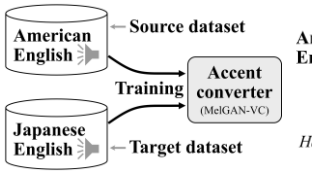

Figure 1: Learning converter

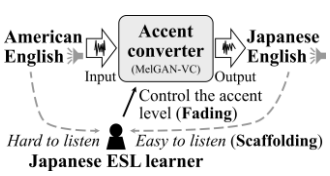

Figure 2: A proposed system

† Department of Information Science, Faculty of Science and Technology, Kochi University

In this paper, we conducted a preliminary experiment that confirmed the clarity of Japanese-accented English. In addition, we confirmed MelGAN-VC[4] can convert from American English voices to Japanese-accented voices for Japanese ESL (English as a Second Language) learners. Because of this, we trained MelGAN-VC using American English as a source dataset and Japanese-accented as a target dataset as shown in Figure 1. In our final project, we are planning to develop an English listening application, that utilizes the gradual change of the level of the accents for the effect of fadable scaffolding. The outline of our project is depicted in Figure 2.

2. Related Work

2.1. SautiLearn

The purpose of SautiLearn[5] was to convert from American English accents to Nigerian accents for listeners who are still not familiar with listening to American English using two methods: MelGAN-VC and FAC-via-PPG[6]. As a result, MelGAN-VC converted a speaker's style instead of converting accents while FAC-via-PPG could accurately convert accents.

However, the previous studies have not been able to establish accent conversion by MelGAN-VC is impossible. In that research, MelGAN-VC used data from CMU-ARCTIC[8] as a source dataset and L2-ARCTIC[7] as a target dataset. When using these datasets that include various characteristics such as speaker's style and accents, style transfer occurred. However, if that dataset is an ideal dataset that only includes differences in accents, accent conversion will occur without audio style transfer. Besides, FAC-via-PPG which uses phonetic posteriorgram sometimes causes misrecognition of phonemes. That means if misrecognition frequently occurs by the noise of audio data, the content of utterance randomly differs. In this situation, MelGAN-VC still has strong potential as a model of accent conversion because MelGAN-VC, a kind of GANs directly uses audio data without recognition. That means this model is robust to noise.

2.2. Synthesized speech

For properly using MelGAN-VC, an ideal dataset that includes only differences in accents is necessary. For this requirement, synthesized speech, TTS voices are useful for making an ideal dataset. Amazon.com provides Amazon Web Services (AWS) with Amazon Polly, TTS services using neural networks. There are functions of accents in Amazon Polly. In a rule of Amazon Polly, accents are clearly defined as phoneme replacements that depend on a speaker's mother tongue.

3. Method

To train MelGAN-VC which can convert from American English voices to Japanese-accented voices, we decided to use Amazon

Polly. In this research, we defined Japanese-accented speech and American-accented speech as follows:

- The Japanese-accented speech is Japanese speech using a command of 'lang:xml lang = "en-US"', which means reading a sentence with an American English accent.
- The American-accented speech is an American speaker's voice in Amazon Polly.

However, before using Amazon Polly for obtaining the ideal dataset, we followed the 4 steps below:

1. We adjust Japanese-accented speech with a similar duration to an American one.
2. We confirm that Japanese-accented English speech is more clear than American one.
3. We make a Japanese-accented dataset for training MelGAN-VC.
4. We convert from American English to Japanese-accented English using MelGAN-VC.

In the 1st step, Amazon Polly's Japanese-accented English speech is mostly longer than American English. For this reason, the speed-adjusting method is necessary.

In the 2nd step, we must confirm that Japanese-accented English speeches are more clear than American ones. It is because if a target dataset is not easy-to-listen-to-speech, GANs cannot convert from an American English speech to an easy-to-listen-to Japanese-accented speech. Therefore, we decided to conduct an evaluation experiment. In this experiment, we use neural TTS voice avatars that use deep neural networks to produce more natural pronunciation than normal TTS an American male, Matthew, and a Japanese male, Takumi.

In the 3rd step, we used CMU-ARCTIC prompts to make the Japanese-accented and American datasets.

In the 4th step, we confirmed the accuracy of the trained model using an acoustic model to identify phonemes from speech and Levenshtein distance. The Levenshtein distance is the minimum number of operations required to make one string the same by three operations of replacement, insertion, and deletion.

Adjusting-method: For adjusting the length of sounds, we adopted a method that uses silent intervals and speed adjustment. Some American speech and a Japanese-accented speech that has identical utterance content compared to an American one are shown in Figure 3. In these speeches, an American speech is 3.29 seconds while a Japanese-accented one is 4.34 seconds. These modelized speech adjusted by a method is shown in Figure 4.

In Figure 4, $P_{A1\sim A4}$ and $P_{B1\sim B5}$ are different voice intervals. In addition, there are minute voice intervals between P_{A1} and P_{A2} such as a phoneme and another phoneme with not-adjusted audio. In these methods, first, we conducted speech segment detection using deep neural networks. Second, we also directly detected voice intervals and minute silence in more detail using audio information in each of the speech segments. Third, we accelerated minute voice intervals by 1.5 times using time-stretching and removing minute silence to match the seconds of American English.

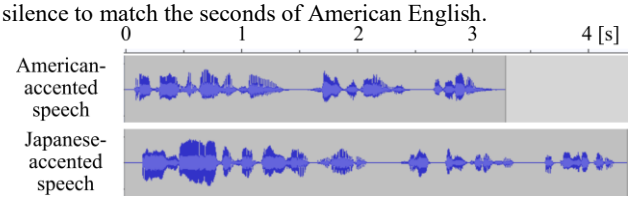

Figure 3: American and Japanese sample audios.

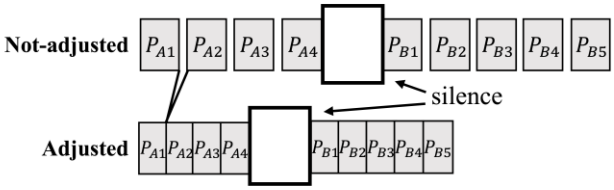

Figure 4: A adopted adjusting-method.

4. Evaluation experiment to confirm clarity of Japanese-accented English

We evaluated the clarity of Japanese-accented speeches compared to American ones. To achieve this purpose, we made English questions.

4.1. Participants

An experiment was conducted in our university between April 20th and 21st, 2023 on university students in two laboratories. We sent each of the students a link to participate in the experiment. After clicking the link, a student can access questions in Google Forms. During the experiment, we used an audio file to announce to solve questions while video conferencing using Microsoft Teams.

In Total, 8 Japanese students participated in the experiment. Table 1 contains information on participants.

Table 1: Information on participants

Participant ID	CEFR
a	A2
b	B1
c	A2
d	Not good
e	A2
f	A2
g	B1
h	Not good

4.2. Design

To confirm the clarity of English, which is independent of a person's listening skills, we did not use standardized tests such as TOEFL. Instead of these tests, for this experiment, we made questions to clear grammatically important points:

- Incomplete sentence: 24 questions
- Questionnaire: 2 questions

using Japanese male Takumi's Japanese-accented speech and American male Matthew's which are produced by Amazon Polly. We used 1.5 times American English (V_{us150}), 1.25 times American English (V_{us125}), and 1 time American English (V_{us100}), as well as, Japanese-accented English: V_{ja150}, V_{ja125}, and V_{ja100} in this evaluation experiment. In addition, we used the following 2 tricks:

1. All questions are randomly generated by a large language model, ChatGPT to avoid making deliberate questions.
2. All question speeches were set to be shorter than the American ones. This is because of the fact that a slower speech is easier to understand compared to a faster one regardless of the influence of the accent. In the end, Japanese-accented speeches were shorter than American ones between 0.0 and 2.05 seconds. Besides, the mean duration difference is 0.26 seconds.

Questions of incomplete sentences: A task of an incomplete sentence consists of short (1 sentence) 12 questions and long (2~7 sentences) 12 questions for confirming the clarity of grammatically important phonemes. In detail, each 12 questions consist of 4 questions in each of 3 sections:

- distinguishing past tense and present tense
- distinguishing positive and negative forms i.e. especially whether or not *'t* and *n't* is attached.
- distinguishing interrogative words: *what*, *when*, *why*, and *how*.

Samples of these types of questions are shown in Table 2.

Table 2: Sample questions

Type	Question
Distinguishing past tense and present tense	Yesterday, I ***(clean/cleaned)*** the entire house for hours, and then I collapsed on the couch because I was so exhausted.
Distinguishing positive and negative forms	The company ***(did/didn't)*** meet its quarterly revenue targets and had to make some difficult decisions about cutting costs and restructuring the business.
Distinguishing interrogative words	***(What/When)*** is the deadline for this project and how much progress have we made so far?

All types of question-right answers do not guarantee the grammatical correctness of the complete sentence. Moreover, all types of question-answer choices have 5 choices, for example, the type of distinguishing past tense and present tense in Table 2, *'clean'*, *think 'clean'*, *don't know*, *think 'cleaned'*, *'cleaned'*. In these answers, if *'clean'* is the right answer, *'clean'* and *think 'clean'* are both correct while *don't know*, *think 'cleaned'*, and *'cleaned'* are wrong.

Questionnaire: A section of the questionnaire consists of these questions:

Q1. Listen to English voices V_{us075} or V_{us100} and V_{ja100}, and evaluate how well you understood English in V_{ja100} compared to V_{us075} or V_{us100} on a scale of 5.

Q2. Listening to Japanese-accented English voice, please write your impressions from various perspectives such as ease of listening and differences from native voices.

In this section, we confirm that Japanese voices are subjectively more comprehensible than American voices.

4.3. Result

Questions of incomplete sentences: The result of questions the incomplete sentence is shown in Table 3.

From Table 3, 7 of 8 subjects with Japanese-accented speech scores are higher than American-accented ones. However, an appropriate statistical test technique for this data, the unpaired t-test showed no significant difference. Although, if there are more subjects, this test may result in significant differences.

Questionnaire: Results of Q1 in the questionnaires are shown in Figure 5. Additionally, as a result of Q2, there were opinions such as difficult to hear due to noise, easy to understand words, and difficult to understand as a sentence. However, in conclusion, there is a point to easy to hear and this speech can be recognized as English for Japanese considering the results of the experiment.

Table 3: Average correct answer rate of questions of incomplete sentences.

Participant ID	Average correct answer rate	
	$V_{ja(150,125,100)}$	$V_{us(150,125,100)}$
a	**88.2%**	72.2%
b	**88.2%**	77.8%
c	**70.6%**	66.7%
d	82.4%	**83.3%**
e	**66.7%**	41.2%
f	**88.9%**	58.8%
g	**66.7%**	58.8%
h	**50.0%**	35.3%

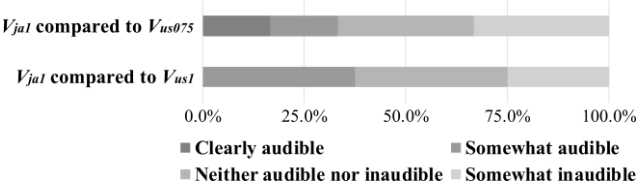

Figure 5:Answers to Q1 in the questionnaire.

5. Experiment to confirm accent conversion using MelGAN-VC

After completing the evaluation experiment, we converted from American-English speech to Japanese-accented speech. For using MelGAN-VC as the converter, we made American-accented speeches and an adjusted-Japanese-accented speeches dataset using CMU-ARCTIC which consists of phoneme-balanced sentences. The length of the American-accented dataset and the Japanese-accented dataset speech, which ranged from 0.02 to 3.24 seconds, was improved to 0.00 to 1.50 seconds by the adjustment method. The mean duration difference was also improved from 1.56 to 0.26 seconds.

5.1. Design

To check if the conversion was successful, we conducted an experiment using Harvard sentences[9] which is a phoneme-balanced sentences dataset composed of 720 sentences, Allosaurus[10] which is an acoustic model, and a similarity algorithm based on Levenshtein distance.

Allosaurus: Allosaurus is a multilingual acoustic model of PER (Phone Error Rate) 20.5% in English. However, in this experiment, we used model eng2102 in Allosaurus which is language-dependent in an English model because multilingual acoustic models recognized minute noise as phonemes.

In this experiment, we used the following data:

P_{conv} : converted Harvard sentences of Matthew's audio of phoneme strings using MelGAN-VC that converted from Matthew to Takumi.

$P_{ideal\text{-}tgt}$: Harvard sentences adjusted audio's phoneme strings which are produced by Takumi.

P_{src} : Harvard sentence audio's phoneme strings which are produced by Matthew.

Using these data, we converted from audio to phoneme string using Allosaurus. In addition, we compared the similarity in Levenshtein distance between P_{conv} and $P_{ideal\text{-}tgt}$ (P_{conv}-$P_{ideal\text{-}tgt}$) as well as P_{conve} and P_{src} (P_{conv}-P_{src}). If P_{conv}-$P_{ideal\text{-}tgt}$ is better than P_{conv}-P_{src} as a result, it can be argued that the conversion is successful from American-accented English to Japanese-accented

English. It is because, on the rule of Amazon Polly, accents are defined as phonetical replacements. For this reason, confirming a change of phonemes means a change of accents.

5.2. Result

This experiment's result is shown in Table 3. Additionally, 2 histograms of P_{conv}-P_{src} and P_{conv}-$P_{ideal\text{-}tgt}$ are shown in Figure 6.

Table 4: phoneme similarity

Data	Mean	Max	Min
P_{conv}-P_{src}	50.1%	73.0%	13.3%
P_{conv}-$P_{ideal\text{-}tgt}$	46.2%	62.9%	7.7%

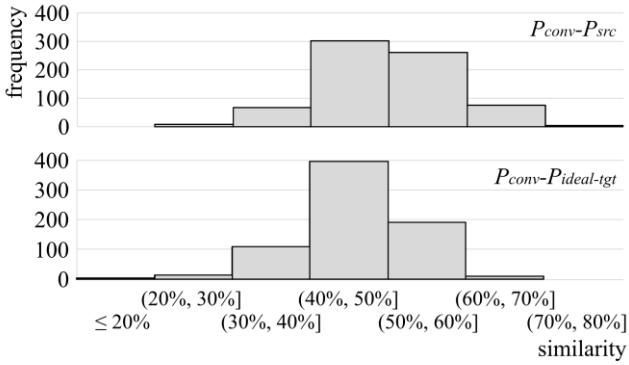

Figure 6: Histogram of similarity of phoneme strings.

As a result, we could not assert MelGAN-VC can completely convert from American English to Japanese-accented English because the mean of P_{conv}-P_{src} is higher than the mean of P_{conv}-$P_{ideal\text{-}tgt}$. However, 2 histograms show the relationship that P_{conv} is related to $P_{ideal\text{-}tgt}$ and P_{src}. Therefore, we investigated the magnitude relationship of the similarity of individual data. From this experiment, the results were as follows:

Case 1: P_{conv}-$P_{ideal\text{-}tgt}$ is higher than P_{conv}-P_{src}. 364 cases.

Case 2: P_{conv}-$P_{ideal\text{-}tgt}$ is lower than P_{conv}-P_{src}. 320 cases.

Case 3: P_{conv}-$P_{ideal\text{-}tgt}$ and P_{conv}-P_{src} are equal. 20 cases.

From this information, we confirmed there are more Case 1 than Case 2. Therefore, we will claim more than half of English audio can be converted to Japanese-accented English using MelGAN-VC.

6. Conclusion and Future Work

In conclusion, we confirmed that Japanese-accented speech is more clear than American one regarding grammatically significant phonemes. In addition, we trained MelGAN-VC using American English and Japanese-accented English in Amazon Polly can convert from some English speech to Japanese-accented speech. However, this experiment to confirm changing accents was not properly working. We expect this cause is attributed to the complexity of the rule of accents in Amazon Polly which frequently causes transformations of phonemes.

From these experiments, future work is necessary to use datasets that include a more simplified Japanese accent than a Japanese accent in the rule of Amazon Polly. Additionally, we must use an algorithm that detects phonemes and accurately aligns the length of one voice with another in more detail. Using these methods, we will make an ideal dataset for MelGAN-VC which can convert from American accents English to Japanese-accented English.

Acknowledgments: This work was supported by JSPS KAKENHI Grant Number 23K02660.

References

[1] Safitri, M. et al. "The correlation between students' reading and listening score in a standardized test of TOEFL." *Journal Penelitian dan Pengkajian Ilmu Pendidikan*: e-Saintika 5.2, pp.127-138, 2021.

[2] Jackson, S. L. et al, and Soloway, E. "The design of guided learner-adaptable scaffolding in interactive learning environments." *Proceedings of the SIGCHI conference on Human factors in computing systems,* pp.187-194, 1998.

[3] Mori, K., and Miyoshi, Y. "Consideration of speech conversion methods using MelGAN-VC for second language learners." *Proceedings of the 97th Conference of SIG-SLUD*. The Japanese Society for Artificial Intelligence, pp.1-4, 2023. (in Japanese)

[4] Pasini, M. "MelGAN-VC: Voice conversion and audio style transfer on arbitrarily long samples using spectrograms." *arXiv preprint arXiv:1910.03713*, 2019.

[5] Afonja, T. et al. "SautiLearn: Improving online learning experience with accent translation." *IDRC・CRDI*, pp.1-11, 2021.

[6] Zhao, G. et al. "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams." *Proceedings of Interspeech 2019*, pp.2843-2847, 2019.

[7] Zhao, G. et al. "L2-ARCTIC: A non-native English speech corpus." *Interspeech*, pp. 2783-2787. 2018

[8] Kominek, J., and Black, A. W. "The CMU Arctic speech databases." *Fifth ISCA workshop on speech synthesis*, pp. 223-224, 2004.

[9] Rothauser, E. H. "IEEE recommended practice for speech quality measurements." *IEEE Transactions on Audio and Electroacoustics* 17.3, pp.225-246, 1969.

[10] Li, Xinjian, et al. "Universal phone recognition with a multilingual allophone system." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* IEEE, pp. 8249-8253, 2020.