PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

A study on English speech acclimatization based on accent conversion for non-native speaker

Yutao ZHANG[(1)], Takuro SASAKI[(1)], Yukoh WAKABAYASHI[(2)], Takahiro FUKUMORI[(1)], Takanobu NISHIURA[(1)]

[(1)]Ritsumeikan University, Japan, gr0320fr@ed.ritsumei.ac.jp, is0215kk@ed.ritsumei.ac.jp, fukumori@fc.ritsumei.ac.jp, nishiura@is.ristumei.ac.jp

[(2)]Tokyo Metropolitan University, Japan / Ritsumeikan University, Japan, wakayoko@fc.ritsumei.ac.jp

ABSTRACT

It is considered that listening and understanding English speech of native speakers are hard tasks for many Japanese. Therefore, we carry out a research on converting the English speech spoken by native speakers based on emphasis of the feature which is familiar to listening to Japanese. We believe the converted speech could be understood more easily by Japanese listeners for English education. We called the research English speech acclimatization. In this paper, we proposed an accent conversion method as an English speech acclimatization method. Specifically, we convert the accent of native speakers into that of Japanese speakers. The acoustic features are time stretched to match the same phoneme duration with the English speech of Japanese speaker. Then the pitch change in English speech of native speaker is replaced with the that in English speech of Japanese speaker. Experiments are conducted to evaluate the performance of proposed method based on the English speech before and after conversion. To confirm the effectiveness on words recognition and comprehension of the sentence, subjective evaluation experiments are conducted respectively. As the result of the evaluation experiments, the effectiveness of proposed method was confirmed.

Keywords: Speech acclimatization, Accent conversion, Pitch, Phoneme duration

1 INTRODUCTION

In times of globalization and internationalization, the opportunities for Japanese people to communicate in English have increased. Nevertheless, because of the difference between speech perception processed by native speakers of difference languages, the communication in English is not an uncomplicated task for many Japanes. According to the various countries' TOEIC average score in terms of 2017 Report on Test Taskers Worldwide provided by ETS (Educational Testing Service), Japanese has lower language proficiency in English [1].

To improve English communication skills for Japanese, we considered a training method based on the idea that the ability of English listening would be improved by carrying on listening training with appropriate English speech according to each level of English listening proficiency. We believed by gradually eliminating the difference between actual native English speech and the perception by non-native speaker, we can expect a good training effect by the sense of accomplishment that trainees can gain easily in this kind of training. In order to achieve this training, it is necessary to create English speech in graded listening difficulty. Meanwhile, the primary goal is to create the English speech with the lowest listening difficulty for Japanese trainee. In other words, the aim of this study is to convert the English speech spoken by native speaker to better understand by Japanese. We called it English speech acclimatization.

In this paper, we propose an accent conversion method as an English speech acclimatization. Accent is variously defined by researches, such as phoneme duration, pitch change, sound pressure, rhythm, etc. In these features, phoneme duration plays an important role in speech intelligibility [2]. Pitch change is an important element in Japanese language perception [3]. Based on these researches, we focus on phoneme duration and pitch change to convert accent in this study. In our method, acoustic features (the pitch change which is vocal cord information and the MFCC (Mel-Frequency Cepstrum Coefficients) which is vocal tract information) in the English speech of native speaker are time-stretched to keep the same phoneme duration with that in the English

speech of Japanese speaker. Then the pitch change in English speech of native speaker is replaced with the that in English speech of Japanese speaker. In the rest of this paper, we will discuss about the differences between the accent in English speech of native speaker and Japanese speaker in Section 2. In Section 3, we introduce the method of accent conversion proposed in this paper. In Section 4, experiments are conducted to evaluate the performance of proposed method based on the English speech before and after conversion.

2 ACCENT IN ENGLISH SPEECH OF NATIVE SPEAKER AND JAPANESE SPEAKER

Accent is a complex expression via the prosodic features such as loudness, pitch and phoneme duration. Generally, English accent belongs to stress accent which mainly depends on loudness, while Japanese accent is considered to belong to pitch accent which mainly depends on pitch change. Figs. 1, 2 and 3 show spectrograms of the English speech of native speaker and the English speech of Japanese speaker. It is clear that the pitch change is different when native speaker and Japanese speak the same English speech. Meanwhile, it is also remarkable that the durations of voiced periods differ from each other. In the English speech of Japanese speaker, there are many pauses (unvoiced parts) which are relatively spaced at regular intervals, while in the English speech of native speaker, there are less pause. In conclusion, the differences between the accent of English and Japanese appear in the same way on English speech of native speaker and Japanese speaker. The English speech of native speaker and Japanese speaker has difference phoneme duration and pitch change. Based on this phenomenon, we believe that it is possible to realize the English speech acclimatization by converting the accent in English speech of native speaker to that in English speech of Japanese speaker which Japanese accustomed.

3 SPEECH ACCLIMATIZATION BASED ON ACCENT CONVERSION

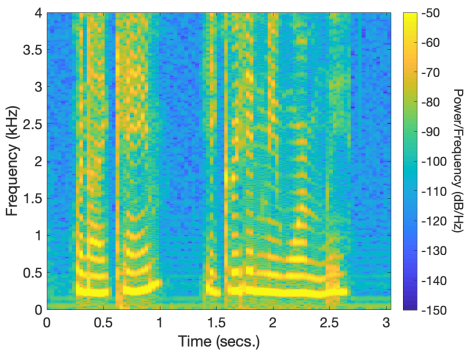
3.1 Overview of accent conversion

As we mentioned before, in order to achieve the English speech acclimatization, the accent in English speech of native speaker is replaced by that in English speech of Japanese speaker. In this paper, the accent refers to phoneme duration and pitch change. The pitch is operated through the operation of the fundamental frequency ($F0$). The outline of the accent conversion algorithm in the experiment is shown in Fig. 4. Here, f_1 indicates the pitch of the native speech and f_2 indicates the pitch of non-native speach. The acoustic features we used in our proposed method are MFCC and pitch. The pitch extraction is performed by SWIPE (Sawtooth Waveform Inspired Pitch Estimator) [4]. Based on SWIPE, the frequency that makes the maximum correlation value between SWIPE kernel and the amplitude spectrum is calculated as the pitch. In order to keep the phoneme duration of native speech and non-native speech consistent, the pitch and MFCC of native speech are time-stretched to match the non-native speech. Then the stretched pitch is converted to mimic the non-native speech. Finally, the converted MFCC and pitch are used to synthesize the converted speech with MLSA filter [5].

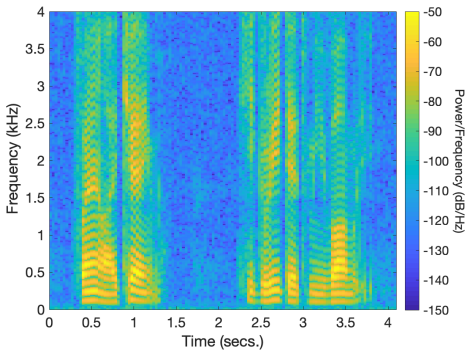
3.2 Feature time stretch

In order to keep the phoneme duration of native speech and non-native speech consistent, the acoustic features (pitch and MFCC) of native speech are linearly time stretched for each phoneme. The duration of each phoneme is extracted with the speech segmentation toolkit of Julius [6]. The example of time stretch is shown in Fig. 5. The time stretched feature can be defined as follows:

$$
\begin{aligned}
m_1'(t) &= (1-b_t)m_1(a_t) + b_t m_1(a_t+1), && (1)\\
a_t &= \lfloor t/k \rfloor, && (2)\\
b_t &= t/k - \lfloor t/k \rfloor, && (3)\\
k &= (T_2-1)/(T_1-1), && (4)
\end{aligned}
$$

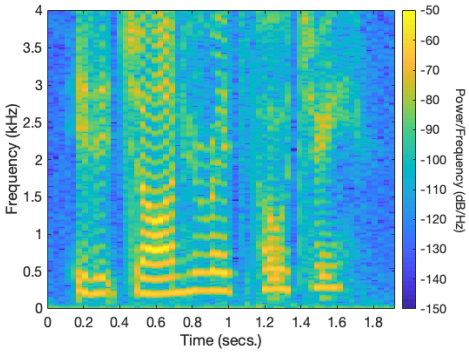


(a) English speech of native speaker.

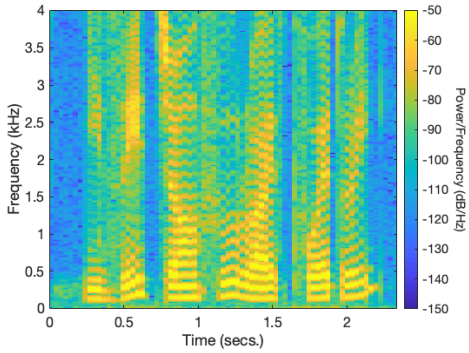


(b) English speech of Japanese speaker.

Figure 1. Spectrograms of the speech “When I came, he greeted me warmly.".

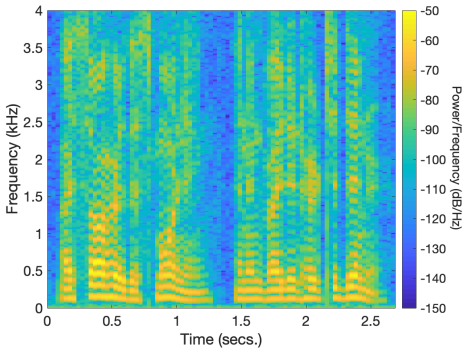


(a) English speech of native speaker.

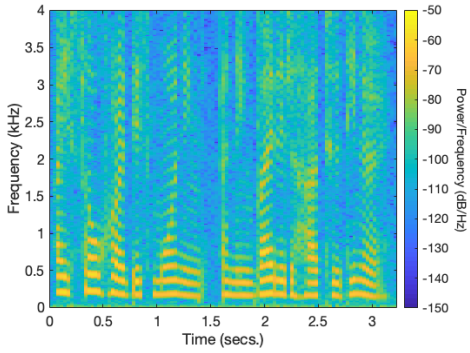


(b) English speech of Japanese speaker.

Figure 2. Spectrograms of the speech “Is it John who writes poetry?".



(a) English speech of native speaker.



(b) English speech of Japanese speaker.

Figure 3. Spectrograms of the speech “I saw her this morning and invited her to dinner.".

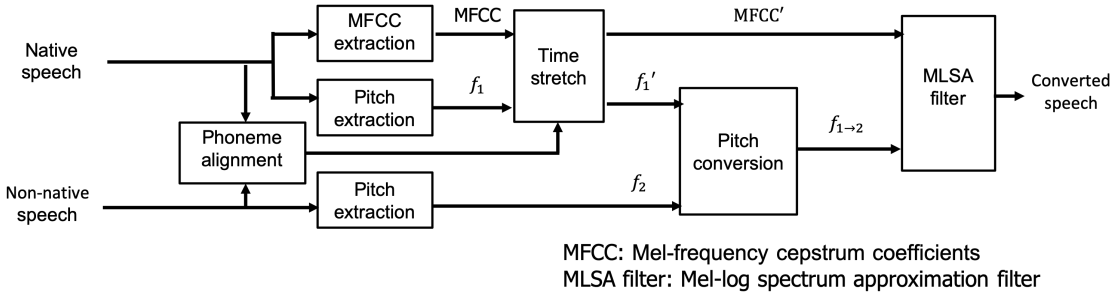


Figure 4. Overview of accent conversion algorithm.

Here, $\lfloor \cdot \rfloor$ represents a floor function. $m_1'(t)$ represents the tth element in the time stretched feature. $m_1'()$ is the time stretched feature linearly interpolated between the samples taken from the $m_1()$ and has the same duration time as $m_2()$. a_t and a_t+1 indicate two samples participates in interpolation, b_t indicates the interpolation ratio and k indicates the stretching factor. T_1 and T_2 correspond to the length of $m_1()$ and $m_2()$ respectively.

3.3 Pitch conversion

The time stretched pitch with the algorithm mentioned in Subsection 3.2 is converted as follows:

$$\log(f_{1\to2}) = \{\log(f_2) - \mu_2\}\frac{\sigma_1'}{\sigma_2} + \mu_1'. \tag{5}$$

Here, μ_1' and σ_1' indicate the time average and standard deviation of $\log(f_1')$ and μ_2 and σ_2 indicate the time average and standard deviation of $\log(f_2)$. f_1' is the pitch which is time sched from f_1 with the algorithm in Subsection 3.1. The converted pitch $f_{1\to2}$, which maintains the change of f_2, can be acquired through the normalization with the mean and variance of f_1'.

Finally, the acclimatized speech is resynthesized from the pitch and MFCC with MLSA filter [5].

4 EVALUATION EXPERIMENT

Two subjective evaluation experiments (Experiment A and B) were carried out to evaluate the effectiveness of English speech acclimatization by accent conversion. In Experiment A, dictation tests were conducted in order to confirm the words recognition. In Experiment B, the multiple-choice tests were conducted based on the response test in the TOEIC to confirm the effectiveness on comprehension of the sentences. In these experiments, the speech before and after accent conversion were compared according to the correct answer rate of the questions. Note that speech distortion caused by speech synthesis interferes fair comparison between native English speech and converted one in terms of comprehension of sentences and word recognition. In order to equalize the effect of the distortion, the native English speech used in these experiments was resynthesized with the acoustic features extracted from the native English speech database.

4.1 Experiment condition

Since it is considered that the recognition performance will be improved when the utterance has been heard several times before, in the experiments, two sets of evaluation speech were used to prevent the subjects from listening to same utterances. The speech before and after conversion was assigned to different sets, meanwhile the number of pre-conversion speech and post-conversion speech in each set was the same. 10 subjects listened to the speech of set 1 and remaining 10 subjects listened to the speech of set 2 with headphone. The sampling frequency of the evaluation speech was 16 kHz, and number of quantization bits was 16 bits.

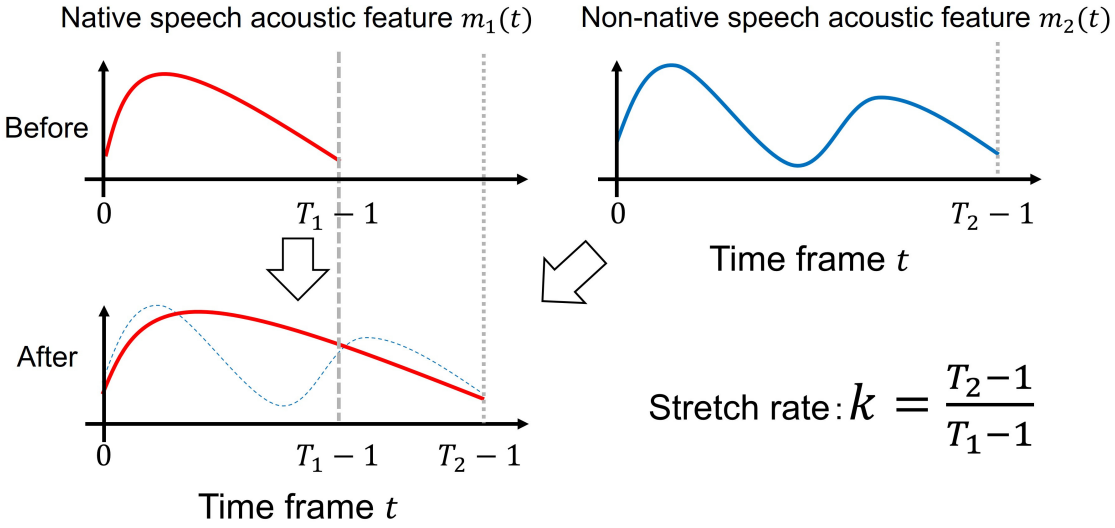


Figure 5. Example of accoustic feature converion by time stretch.

4.1.1 Experiment A

The conditions of Experiment A are shown in Table 1. We used the UME-ERJ database [7] which contains the English speech of Japanese speakers and native speakers. The sentences of the evaluation speech were selected from the database. Subjects were asked to listen to 12 periods of speech and dictate all the words (67 words) they heard. Meanwhile, the inflections of words, such as past tense of verb, incorrectly dictated by subjects would not be judged as mistakes. We also avoided the speech includes names when we chose it from database. The accuracy of dictation was calculated.

Table 1. Condition of Experiment A.

Native speech	Native speaker English speech in UME-ERJ database [7]
Non-native speech	Non-native speaker English speech in UME-ERJ database [7]
Question form	Dictation (12 questions)
Times of listening for each speech	2
Subjects	20 Japanese students

4.1.2 Experiment B

Table 2 shows the experiment conditions of Experiment B. The speech in the response test of TOEIC support material was used as the English speech of native speaker in this experiment. Meanwhile, a Japanese male in 20s was required to speak the same sentences as the English speech of Japanese speaker. We prepared 12 questions in this experiment. In each question, as an example shown in Table 3, 20 subjects were asked to listen to the speech and choose the best response they thought for the speech. The number of choices in each question was 3.

Table 2. Condition of Experiment B.

Native speech	Speech in TOEIC support material [8]
Non-native speech	Speech spoken by a Japanese male in 20s
Question form	Multiple-choice test (12 questions)

Table 3. An example of response test in Experiment B.

Speech	Whose file is this on my desk?
Question	(A) No, that's not my desk. (B) I've been looking for that. Sorry! (C) Mr. Smith called ten minute ago.

4.2 Experiment result

4.2.1 Experiment A

The experiment results are shown in Table 4. In the average of set 1 and set 2, the accuracy of dictations with the pre-conversion speech is 59.6%, and that with the post-conversion speech is 63.7%, showing an improvement of 4.1%. This tendency also can be confirmed in each set of the experiment. The results indicate the effectiveness of accent conversion on the improvement of word recognition.

Table 4. The accuracy with and without accent conversion [%].

		Pre-conversion speech	Post-conversion speech
Experiment A	Set 1	56.8	61.0
	Set 2	62.4	66.5
	Average	59.6	63.7
Experiment B	Set 1	55.0	66.7
	Set 2	51.7	68.3
	Average	53.3	67.5

4.2.2 Experiment B

The experiment results are shown in Table 4. In the average of set 1 and set 2, the improvement of accuracy is 14.2%. Meanwhile, the improvement in set 1 is 11.7%, and the improvement in set 2 is 16.6%. The results show that the accent conversion in proposed method can improve the performance of sentence comprehension.

4.3 Analysis

According to the improvement rate of each experiment, the improvement rate of the experiment A is larger than that in the experiment B. The results indicate that the proposed method is more effective in the sentence comprehension than the word recognition. The reason should be that the proposed method converts accent not just for words, but for the whole sentence. Improving the performance of word recognition is just a part of the advantage of this method.

As the feedback of the subjects who had a high accuracy for the converted speech, the sufficient phoneme duration and acquainted pitch change are useful for understanding the English speech. On the other hand, as the feedback of the subjects who had a high accuracy for the post-conversion speech, there was an opinion that the converted speech is difficult to understand since it sounds like unnatural speech. Therefore, it is necessary

to carry out the study of reducing the feeling of unnatural in post-conversion speech in the future.

5 CONCLUSION

In this paper, we proposed an English speech acclimatization method that converts accent of English speech of native speaker into the accent of Japanese speaker. As the result of the subjective evaluation experiment, the effectiveness of the English speech acclimatization with accent conversion has been confirmed. In the feature, we will focus on the applications such as controlling the conversion rate of the accent, aiming to synthesize the speech suitable for each level of English listening proficiency.

ACKNOWLEDGEMENTS

This work was supported by R-GIRO (Ritsumeikan Global Innovation Research Organization) funded by Ritsumeikan University, JST COI, and JSPS KAKENHI Grant Numbers JP18K19829, JP19H04142.

REFERENCES

[1] Educational Testing Service, "2017 report on test takers worldwide," 2017. https://www.etsglobal.org/About-us/News/2017-Report-on-Test-Takers-Worldwide

[2] K. Tajima, R. Port, and J. Dalby, "Effects of temporal correction on intelligibility of foreign-accented English," Journal of Phonetics, vol. 25, no. 1, pp. 1–24, 1997.

[3] K. Hirose, "Speech Prosody and CALL," Journal of the Phonetic Society of Japan, vol. 9, no. 2, pp. 38–46, 2005.

[4] A. Camacho, and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," The Journal of the Acoustical Society of America, vol. 124, no. 3, pp. 1638–1652, 2008.

[5] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan (Part I: Communications), vol. 66, no. 2, pp. 10–18, 1983.

[6] A. Lee, and T. Kawahara, "Recent development of open-source speech recognition engine julius," Asia-Pacific Signal and Information Processing Association, pp. 131–137, 2009.

[7] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database spoken by Japanese learners," In Reprint of the COCOSDA Workshop, pp. 76–81, 2001.

[8] Y. Mori, "New TOEIC test strategy royal road listening edition [in Japanese]," Asahi Press, Japan, 2008.