Analysis of preferred speaking rate and pause in spoken easy Japanese for non-native listeners

Hafiyan Prafiyanto*, Takashi Nose†, Yuya Chiba‡ and Akinori Ito§

Graduate School of Engineering, Tohoku University,
Aramaki Aza-Aoba 6–6–05, Aoba-ku, Sendai, 980–8579 Japan

(Received 10 April 2017, Accepted for publication 1 August 2017)

Abstract: We investigate the effect of speaking rate and pauses on the perception of spoken Easy Japanese, which is Japanese language with mostly easy words to facilitate understanding by non-native speakers. In this research, we used synthetic speech with various speaking rates, pause positions, and pause lengths to investigate how they correlate with the perception of Easy Japanese for non-native speakers of Japanese. We found that speech rates of 320 and 360 morae per minute are perceived to be close to the ideal speaking rate. Inserting pauses in natural places for Japanese native speakers, based on the dependency relation rule of Japanese, makes sentences easier to listen to for non-native speakers as well, whereas inserting too many pauses makes the sentences hard to listen to.

Keywords: Language for special purposes, Second language acquisition, Educational technology & language


1. INTRODUCTION

Nowadays many foreigners live in Japan, and the number of the foreigners has doubled since 1990 [1], reaching more than 2 million residents. However, many of them do not speak Japanese very fluently, and sometimes they cannot understand public information provided in the Japanese language. One solution is to provide the information using common foreign languages, such as English, Chinese, and Korean. In times of disaster in particular, public information can become vital: when a disaster such as a tsunami or earthquake strikes, an inability to understand information quickly can become a matter of life and death, and at such times it is impractical to provide such information in several languages because of the time needed for translation.

A more practical way to convey information to non-native speakers of Japanese is to use simplified Japanese that has a limited vocabulary of easy words and avoids difficult grammar. “Easy Japanese,” proposed by Sato et al., is one implementation of this idea [2]. It defines the vocabulary and grammatical structures for making Japanese sentences easily understood by foreigners. Easy Japanese was used in written announcements such as posters that were displayed in places of refuge during the Great East Japan Earthquake in 2011, helping many foreigners.

In addition to written announcements, it is often necessary to convey information in spoken language. The radio broadcast is an important way of announcing public information under a disaster, because a battery-powered radio broadcast receiver can be used when a blackout happens. The announcement through the radio broadcast is usually spoken by a human announcer; in addition, use of synthetic speech is also being considered [3,4]. The main advantage of using synthetic speech instead of a human announcer is a more streamlined process from composing the information to broadcasting it. Faster compilation and broadcasting of information is very helpful in times of disasters, where information needs to be updated frequently. Therefore, the use of speech synthesis for Easy Japanese broadcasts is worth considering. For announcements, the temporal properties of the spoken utterances can affect the perception and understanding of the announcement. To ensure that spoken announcements can be understood by the listeners, it is necessary to investigate the effect of those properties on the perception of Easy Japanese announcements.

This research examined prosodic parameters that might affect the perception and intelligibility of spoken an-
nouncements in Easy Japanese for non-native listeners. In this paper, we investigate the effect of speaking rate, pause position, and pause length, hoping to identify an optimum speaking rate, pause position, and pause length. Assessments of listeners’ understanding and preference of the perceived speech were used to judge the optimality.

2. BACKGROUND

2.1. Concept and Design of Easy Japanese

Easy Japanese was proposed by Sato et al. in 1999 following the Hanshin-Awaji Great Earthquake in 1995 [2]. It is a similar idea to Basic English [5]. Easy Japanese is designed to be easier to be understood by foreigners living in Japan. The Easy Japanese Guideline published by the Sociolinguistic Laboratory of Hirosaki University provides some rules for composing information in Easy Japanese [6], such as: (a) The vocabulary should be restricted to the words that appear in the N4 Level of the Japanese Language Proficiency Test (JLPT) or easier, (b) Terms related to the main event, e.g. ‘earthquake,’ should be used even if they are not included in (a), (c) Sentences written in Easy Japanese should be grammatically simple sentences (not compound sentences), (d) Sentences should be short (around 24 morae on average), and (e) Expressions that suggest possibility or supposition (e.g., ‘might be’ or ‘it could happen’) should be avoided.

An example of a normal Japanese sentence and its translation to Easy Japanese is shown in Fig. 1. Considering the sentence in normal Japanese, the meaning of the word ‘teiden’ (blackout) can be easily understood by a native speaker, but it might be a difficult word for a non-native speaker. The structure of this sentence is also difficult; to make it easier to understand, the structure is changed while keeping the intended meaning, using easier words like ‘denki’ (electricity). Also, unnecessary information is deleted to make the sentence shorter.

2.2. Support Systems Developed for Easy Japanese

One problem with converting normal Japanese sentences to Easy Japanese sentences is that it is hard for average native Japanese speakers to judge the easiness of words for non-native speakers, so it is difficult to decide which words need to be converted. To help native speakers writing sentences in Easy Japanese, Nagano and Ito [7] developed the software YANSIS (YAsashii Nihongo SIen System in Japanese, Easy Japanese Support System). YANSIS can point out the parts of the input sentence that do not follow the rules of Easy Japanese, such as words and phrases that might be difficult for non-native speakers. YANSIS also has a function to determine the easiness of a sentence [8]. This function assigns a score of easiness to each input sentence based on sentence structure, difficulty of words, use of foreign words, and use of symbol characters.

2.3. Effect of Prosodic Properties on the Perception of Spoken Language by Non-native Listeners

In past studies, most investigations on the understanding of Japanese language by non-native listeners were done at the segmental level like syllable, mora, and phoneme. For example, there are studies regarding non-native Japanese listeners’ ability to discern difficult phonemes like affricate consonants [9] or special morae, including geminates [10] and elongated vowels [11]. These studies provide important clues on making utterances easier to understand. On the other hand, properties at the suprasegmental level, i.e. prosody, also affects the comprehension of the sentence. Example of such prosodic properties include speaking rate, rhythm [12], pause insertion, fundamental frequency, and intensity [13]. Studies on the effect of prosodic properties on the intelligibility of spoken sentences are still mostly done on native listeners of Japanese, particularly those with impaired hearing due to aging [14]. We can see that the effect of these prosodic factors on the perception of spoken Easy Japanese by non-native listeners has still not been sufficiently investigated.

The effect of prosodic properties on the intelligibility of sentences for non-native speakers has been studied for many other languages. Particularly, speaking rate is a parameter widely believed to be related to intelligibility [15]. For English, a fast speaking rate has been shown to hinder understanding for non-native speakers [16]. In contrast, the benefit of a slow speaking rate on understanding is disputed [17], and appears only in limited circumstances. A speaking rate which is too slow can even hinder understanding, possibly because it disturbs the listener’s concentration [18]. Besides speaking rate, pause is another important prosodic property that has been researched in some languages. In English and many other languages, pauses in appropriate positions can help understanding [19]. However, too many pauses, or pauses that are too long, can also hinder understanding for non-native speakers [17].

Prosodic properties of utterances are, to a certain degree, language dependent. An analysis of several
languages shows that Japanese is a comparatively fast language, with around 7.84 syllables per second in everyday conversation as opposed to 5.22–7.18 syllables per second for many other languages [20]. For non-native Japanese listeners, this may mean that slower Japanese language utterances can be helpful, even more helpful than for other languages. Therefore, the effect of speaking rate on the understanding of Japanese language by non-native listeners needs to be investigated. Furthermore, the appropriate locations of pauses are greatly affected by the relations between the words decided by grammatical rules. Because grammatical rules are language dependent, it is necessary to investigate the effect of pauses in Japanese language as well.

3. SPOKEN EASY JAPANESE FOR NON-NATIVE LISTENERS

The intelligibility of a written announcement depends mostly on the difficulties of its words and grammatical rules. Using Easy Japanese can help keeping these difficulties minimum. When a script written in Easy Japanese is orally announced, additional factors, such as the speaking rate and pauses, will also influence its intelligibility. To deliver spoken announcements in Easy Japanese that can be easily understood, we need to investigate how these factors affect the perception of spoken Easy Japanese sentences.

In this study, we use synthetic speech to investigate preferable prosodic properties. As mentioned in the introduction, speech synthesis enables the information to be delivered faster than relying on human announcers. Also, speech synthesis can be used in places such as small schools that have a network of loudspeakers but do not have staff trained to deliver announcement professionally.

Table 1 summarizes the measures we use in this research. We find the understanding of a sentence using a subjective measure and an objective measure. We use the term “comprehensibility” for the subjective measure and “intelligibility” for the objective measure. Both are common ways to measure understanding [21]. To assess the comprehensibility, the listener assesses his/her own understanding. Meanwhile, dictation is used to assess the intelligibility. Some authors, including Kachru and Smith [22] argue that an ability to write down an utterance shows some capabilities of recognition of a word or another sentence-level element of an utterance, and can be used to measure intelligibility eventhough the listener could not understand the whole meaning. We expect that these measures can be higher when a sentence is spoken slowly (slow speaking rate, more pauses, or longer pauses inserted). However, we expect that a sentence that is too slow might not have much improvement to the intelligibility and comprehensibility compared to a slightly slow sentence, and might even have negative effects. Therefore, we also find listeners’ preference of the speed of the sentence by measuring the “speed adequacy” and the “listenability.” Speed adequacy is measured by asking the listener how close the speed felt compared to the speed he/she feels most appropriate [23], while listenability is measured by asking the listener how easy to listen to the sentence [24]. The listenability is often used as an index of speech quality of synthesized or enhanced speech [25,26].

We conducted three experiments to investigate the optimal speaking rate, pause position, and pause length. Finally, we also conducted an experiment to investigate whether applying all of those optimal conditions would make sentences sound better, by comparing it with the sound at the standard speaking rate. In total, 67 subjects participated. The participants are international students from 8 countries, most (62 participants) came from Asian countries. 23 participants had taken Japanese Language Proficiency Test, while most participants that never took the test reported that they did not use much Japanese in their daily life.

4. INVESTIGATION OF APPROPRIATE PROSODIC PARAMETERS

4.1. Speech Synthesis Based on Hidden Markov Models (HMMs)

We used HTS as a speech synthesizer, which is a widely-used parametric speech synthesis system based on hidden Markov models (HMMs) [27]. The speech synthesized using this method can achieve an intelligibility comparable to natural human speech [28]. Also, this method lets us control the duration of each phoneme, and we use this to control the speaking rate.

Pauses are usually considered to be a prosodic feature, but can be modelled as phonemes in HMM-based speech synthesis. The advantage of treating a pause as a phoneme is the ability to insert pauses by simply adding the pause phoneme. The length of the pause can also be set by explicitly defining the duration of the pause phoneme.

4.2. Speaking Rate Control

To find an appropriate speaking rate, we synthesize
speech with various speaking rates. The speaking rates are controlled by uniformly and linearly converting phoneme durations in the parameter generation process. A survey of broadcasted programs shows that the average speaking rate on Japanese programs ranges from 450 to 570 morae per minute [29]. We assume that an easier speaking rate for understanding by non-native speakers would be slower, so in this research we tested the effect of five speaking rates: 240, 280, 320, 360, and 400 morae per minute.

4.3. Pause Insertion

4.3.1. Pause position

In speech communication, it is known that pause control is very important and pauses inserted in appropriate positions can help a spoken sentence sound more natural and intelligible [19]. In this study, we examined three types of pause insertion position as follows:
(a) **None**: No pause is inserted.
(b) **Dependency**: Pauses are inserted according to the rule described below, illustrated in Fig. 2(a).
(c) **Phrase**: Pauses are inserted between all phrases, illustrated in Fig. 2(b).

A Japanese sentence consists of one or more phrases [30], which are important for determining the positions of pauses. A typical Japanese phrase (bunsetsu) contains one content word and zero or more function words such as particles, suffixes or auxiliary verbs. In Japanese, a pair of phrases can have a dependency relation, with one phrase being a head and another being a dependent. Japanese is a head-final language, which means that the head always comes after the dependent, but not always immediately after it. The distance between the head and the dependent in the unit of phrase is called the dependency relation distance. If this distance is more than one, which means there are one or more phrases sandwiched between the head and the dependent, it is likely that a pause is inserted immediately after the dependent in natural Japanese speech [31].

Even though the dependency-based pause insertion points are natural for a native speaker of Japanese, more pauses might help non-native speakers of Japanese to understand the sentence better. The Easy Japanese Guideline suggests inserting pauses between each phrase boundary [6].

4.3.2. Pause length

For a silent region within speech to be recognized as a pause, the length must be at least around 200 ms [19]. However, the best pause length for non-native speakers is not known, and it might be longer. In this research, we examined pause lengths between 200 and 800 ms.

5. EXPERIMENTS

5.1. Experimental Procedure

In our research, we conducted three listening tests. The first experiment examined the effect of speaking rate, the second examined the effect of the positions of pauses within the sentence, and the third examined the effect of pause length.

Each experiment consisted of two parts, the subjective evaluation part and the dictation part. In the subjective evaluation part, the subjects of the experiment listened to fifteen Easy Japanese sentences with either the speaking rate, pause insertion rule or pause length controlled, and then rated the listenability score and the comprehensibility score of each sentence. The listenability score was rated by asking how easy to listen to the sentence, with the score ranging from 1 (very hard) to 5 (very easy). The comprehensibility score was rated by asking how much of the meaning of the sentence is understood, with the score ranging from 1 (does not understand at all) to 5 (completely understand). Finally, the subjects also rated how fast the sentence felt on a scale from 1 (very slow) through 3 (just right) to 5 (very fast). Using the rated speed, we defined a speed adequacy score, which showed how close the speed felt compared to the most appropriate speed. Because we defined option 3 as “just right” for the perceived speed question, the perceived speed of 3 is defined as the most appropriate speed, with the speed adequacy of 3. The perceived speed of 2 and 4 have the speed adequacy of 2, while the perceived speed of 1 and 5 have the speed adequacy of 1.

We objectively measure intelligibility in the dictation part. In this part, the subjects listened to eight Easy Japanese sentences with various speaking rate and pause conditions, then typed each sentence. From the subjects’ answers, we defined a dictation score for each sentence. The dictation score was calculated using the Levenshtein distance between the spoken sentence and the subject’s answer, inverted and linearly normalized between 0 (maximum distance, lowest score) and 1 (zero distance, perfect score).

Let $s_{\text{ans}}$ and $s_{\text{ref}}$ be the subject’s dictation and the reference sentence, respectively, written in hiragana. Let
As follows:

\[ S_{\text{dict}} = \frac{\max(n_{\text{ans}}, n_{\text{ref}}) - d_1(n_{\text{ans}}, s_{\text{ref}})}{\max(n_{\text{ans}}, n_{\text{ref}})} \]

Here, \( d_1 \) is the Levenshtein distance.

Table 2 shows the sentences used in this experiment. Sentences 1 to 15 were used in the subjective evaluation part of the experiments, while the dictation part of the experiments used sentences 1 to 8 only. The slashes (/) indicate the positions inserted according to both the dependency rule and the phrase rule.

After conducting the experiments regarding speaking rate, pause position, and pause length, we tried to determine the best condition for each factor. We then conducted an experiment to investigate whether applying all of those conditions would make sentences sound better, by comparing it with the sound at the standard speaking rate. Participants were asked to listen to sentences with standard prosodic parameters and sentences with tuned prosodic parameters, then rate which ones were more subjectively listenable, intelligible, and natural.

5.2. Result for Speaking Rate

In the experiment regarding speaking rate, the subjects listened to Easy Japanese sentences with five speaking rates (240, 280, 320, 360 and 400 morae per minute). The sentences contained no pauses. The subjects were 21 international students at a Japanese university who had lived in Japan for 4 to 70 months (average: 2 years). 19 participants came from Asian countries (China and Indonesia), while 2 participants came from Latin America. 7 participants had JLPT qualifications.

Figure 3(a) shows the relation between the speaking rate and the average speed adequacy. The graph shows that the speaking rates of 320 and 360 morae per minute have the highest speed adequacies. Using one-way ANOVA, the speaking rate was shown to be a statistically significant \((p < 0.0001)\) factor influencing the speed adequacy. Using Tukey’s test, we found that the average speed qualities of both 320 and 360 morae per minute were significantly higher than 240 morae per minute \((p < 0.0001)\) and 280 morae per minute \((p < 0.001)\).

Figures 3(b)–3(d) show the relation between the speaking rate and, respectively, the average listenability, the comprehensibility, and the dictation score. Using ANOVA, the differences between each rate for those scores were not statistically significant at the 5% significance level, suggesting that the scores were hardly affected by the speaking rate.

Based on the results of this experiment, we conclude that the speaking rates of 320 and 360 morae per minute are significantly better. Furthermore, because the speaking rate of 360 morae per minute can deliver information faster, it is preferable to 320 morae per minute for efficiency of information communication. Therefore, for the next ex-
experiments regarding pauses, we determined the effect of pause insertion in sentences spoken at the rate of 360 morae per minute.

5.3. Result for Pause Position

In the experiment regarding pause position, the subjects listened to Easy Japanese sentences with the three pause position rules (none, dependency, phrase). The pause length was fixed to 500 ms. The subjects were 18 international students at a Japanese university who were not a subset of the participants in the previous experiments, although they too had lived in Japan for 4 to 70 months (average: 2 years). 16 participants came from Asian countries (China, Indonesia, Korea), while 3 participants came from Europe. 6 participants had JLPT qualifications.

Figure 4 shows the average scores for each pause position rule: (a) speed adequacy, (b) listenability, (c) comprehensibility, and (d) dictation score (***: $p < 0.001$, ****: $p < 0.0001$).

**Figure 4** Average scores for each pause position rule: (a) speed adequacy, (b) listenability, (c) comprehensibility, and (d) dictation score (*: $p < 0.05$, ***: $p < 0.001$, ****: $p < 0.0001$).

Based on the results of this experiment, we conclude that the pause position based on the dependency rule has significantly higher listenability. Therefore, in the next
experiment, we determined the effect of pause length by inserting pauses at the positions determined by this rule.

5.4. Result for Pause Length

In the experiment regarding pause length, the subjects listened to Easy Japanese sentences with five conditions of pause length (200, 350, 500, 650, and 800 ms). The subjects were 17 international students at a Japanese university who had lived in Japan for 1 to 77 months (average: 2 years). All participants came from Asian countries (China, Indonesia, Korea). 7 participants had JLPT qualifications.

Figure 5(a) shows the relation between the speaking rate and the average speed adequacy. The graph shows that the pause lengths of 200–500 ms had higher speed adequacy than those of 650 and 800 ms. Using one-way ANOVA, pause length was shown to be a statistically significant ($p < 0.05$) factor influencing the speed adequacy. However, using Tukey’s test, we found no significant difference at the 5% significance level between the average speed adequacies of any scores. This suggests that while pause length influenced how fast a sentence sounds, we cannot say with confidence which pause length was better.

Figures 5(b)–5(d) show the relation between the pause length and, respectively, the average listenability, the comprehensibility, and the dictation score. Using ANOVA, the differences between each pause length for those scores were not statistically significant at the 5% significance level. This suggests that those scores were hardly affected by the pause length.

Based on the results of this experiment, we conclude that pause length has no significant effect on listenability or comprehensibility. It can be argued that the pause length of 200 ms is the best option, because it allows information to be transferred faster.

5.5. Comparison of Standard and Tuned Speech

In the experiment comparing standard and tuned speech, the subjects listened to Easy Japanese sentences with standard parameters of the speech synthesizer (speaking rate of 417 morae per minute, no pauses) and tuned parameters (speaking rate of 360 morae per minute, pause position based on the dependency relation rule, pause length of 200 ms). The subjects were 10 international students at a Japanese university who had lived in Japan for 2 years on average. All came from Asian countries (China and Indonesia). 3 participants had JLPT qualifications.

Figure 6 shows the comparison between the standard speech and the tuned speech. The tuned speech was found subjectively to be more listenable and intelligible, but less natural than the standard speech. Using the population proportion test, the differences were found to be significant ($p < 0.05$). The speech rate of 360 morae per minute is rather slow and unlike usual speeches in broadcasts; this might explain the decrease in the naturalness.

6. DISCUSSION

We mentioned in the introduction that one important reason for this research is to make spoken announcements such as radio broadcasts in Easy Japanese easy to listen and understand for non-native speakers. In this study, we measured understanding using comprehensibility and intelligibility. We measured preference using speed adequacy and listenability. On the experiments measuring the effect of various prosodic conditions on the understanding and preference, we found that generally, the understanding was not affected much, but the preference was affected: a speaking rate of 360 morae per minute is perceived to be close to the ideal speed by non-native speakers and has high listenability when pauses are inserted in appropriate
positions based on the dependency relation rule. In this section, we will discuss about the validity of this measurement, and other things that need to be considered when applying our findings.

The methods used to measure the comprehensibility (self assessment) and the intelligibility (dictation) in this study has some caveats. Self assessment cannot catch genuine misunderstanding on the listener’s part, so it might overestimate the understanding. Dictation requires memorization skill and writing skill in addition to understanding, so it might underestimate the understanding. However, we found that the correlations between the comprehensibility and the intelligibility ranged from moderate to strong (speaking rate: \( r = 0.52 \), pause position: \( r = 0.81 \), pause length: \( r = 0.78 \)), showing that they were reliable to a certain degree. Although we expected that the understanding would be higher for slower conditions based on observations in daily life, the fact that no differences were found was not unprecedented, and agrees with some previous studies [17,18].

We found that the speed adequacy was the only measure significantly affected by all of the conditions. However, the differences in the perceived did not translate to the difference in the intelligibility and the comprehensibility. This suggests that even for conditions thought to be faster than ideal, the subjects can still comprehend parts of the sentence or get the dictation of easier words right. The speaking rate of 360 morae per minute was perceived to be close to the ideal speed by non-native speakers. This rate is substantially slower than the average speaking rate usually found in programs for native speakers. This rate is substantially slower than the average speaking rate usually found in programs for native speakers, which ranges from 450 to 570 morae per minute [29]. An important point about radio broadcasts is that they are listened to by both native speakers and non-native speakers, who have different preferences for speaking style. Before applying the results of this research, it is important to consider the best way of speaking announcements from the viewpoint of universal design.

7. CONCLUSIONS

We investigated the effects of prosodic properties (speaking rate, pause position and pause length) on the perception of speech. We used synthetic speech with various conditions to investigate how they correlate with the intelligibility and listenability of spoken Easy Japanese. We found that a speaking rate of 360 morae per minute with 200 ms pauses to be close to the ideal speaking speed. It is also more appropriate to insert pauses at appropriate natural positions for native speakers, based on the dependency relation rule of Japanese language, as opposed to inserting pauses between every phrases. Speech under these conditions was found to be more listenable and intelligible than speech at the standard speaking rate.

ACKNOWLEDGEMENT

Part of this work is supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number JP26284069 and JP16K13253. We also give thanks to Prof. Kazuyuki Sato of Hirosaki University for useful discussions.

REFERENCES


