Relation between Pause and Text Structure in Japanese Story Telling

—Towards a more live text-to-speech synthesis—

Akihito AOYAMA
Science University of Tokyo
aa@mt.is.noda.sut.ac.jp

Kumiko TANAKA-Ishii
The University of Tokyo
kumiko@ipl.t.u-tokyo.ac.jp

Masayuki TAKEDA
Science University of Tokyo
takeda@mt.is.noda.sut.ac.jp

Abstract

With the objective to make TTS more lively, we present our results of analysis on Japanese speech data of story telling, focusing on the pause lengths. We first see that the longer pause is taken between the larger text elements. Also, pause lengths become shorter than average when two phrases or sentences are tightly related. We also found that the emphasis is expressed by taking the longer pause.

keywords: pause, live text-to-speech, text analysis

1 Introduction

Text to speech software (TTS in the following) has nowadays become more popular. Not only it is used in the research area such as talking robots, it is now starting to be used more in daily life, to let the user know the status of electric apparatus, or to let the foreign language learners know the pronunciation of his target language. The technology of TTS has indeed become mature in the sense that TTS can read sentences without making mistakes of where to pause, how to read a sentence using correct phonemes with average sort of intonation.

However, the technology is far from mature in the sense of liveliness, that is reading out in more natural tone according to the context. Nass and Reeves[3] indicate that when TTS reads out a text, it is important for the human listeners that the emotional tone and the context accord. For example, listeners cannot accept a sad text read out in a gay tone as being natural.

In fact, TTS could be controlled to sound lively using *commands*, that are equipped in most of the TTS software. Users may change pitch, stress and pause locally by embedding these commands into text changing parameters. Though, here arises the problem. Where should we put these tags? How should we differ the parameters?

The scientific research around this question is in its infancy. As for pitch, Pan[1] examined how context affects intonations with speech data of doctor/patient diagnosis. Fry[4] studied a conversation in Japanese and indicated that accent is put when new topic arises. As for pause, however, there are even less previous works. Sugitou[5], in her textbook for reading, indicates how pause affects the listeners' impression. For example, if longer pause is taken after a sentence, the listener feels that the sentence was read out slower than usual, although the sentence was in fact read out at the same speed. She reveals how pause could be used to sound differently.

Strongly motivated by her work, we analyze pause in a speech data, so that TTS could sound more lively by varying places and length of pauses. First, we show how people are sensitive to pause length by small experiments. Then we revise Sugitou's work to clarify what law we look for before we go through data in our experiments. Then, we examine speech data to verify whether the actual speech data really follows the law.

2 The Importance of Pause for Human Listeners

Before we go on to the analysis of pause length, we present our small experiment that shows how pause length is related to the good impression of speech.

We prepared the following two speech data:

Original-A,B A small parts of a story read up by a professional story teller (more explained in §1) (about 30secs each).

Modified-C,D The same data processed so that the data has the unique length of pause between any successive two phrases. Note that data A is processed into C and data B into D.

Note that audio processing was carefully performed so that the results does not contain any additional noise to be heard compared with the original.

10 examinees were asked to listen to the two of the above four data. All examinees are asked to score the impression from the following viewpoints:

Ease of listening (easy(5) difficult(1))

Comprehension of the context (understandable(5) not understandable(1))

Liveliness (live(5) dull(1))

Imagination (more (5) less(1))

Naturalness (like human(5) like machine(1))

Higher scores mean that examinees have the better impression of the data.

Among 10 examinees, 5 examinees (group 1) were asked to listen to the data in the order of original (Original-A) then processed data (Modified-D) whereas the other 5 (group 2) are asked to listen to the data in the reverse order, first processed (Modified-C) and then original (Original-B).

Figure 1 shows the results. The vertical axis denotes the average score of examinees and the horizontal axis denotes each viewpoints. Each line shows the average impression of data by an group (thus denoted such as Natural-A-Group1). When the lines is located higher in the figure, the data is perceived better.

We could clearly see that the line for data A and B is located higher than that of data C and D. The audience's reaction was so clear that they could sense the pause even for the short speech data of 30 msecs. Note that nothing else than pause is processed when we modified A and B into C and D. From this result, we may say that proper pause length plays an important role to gain the better impression even for the short data of 30 sec length.

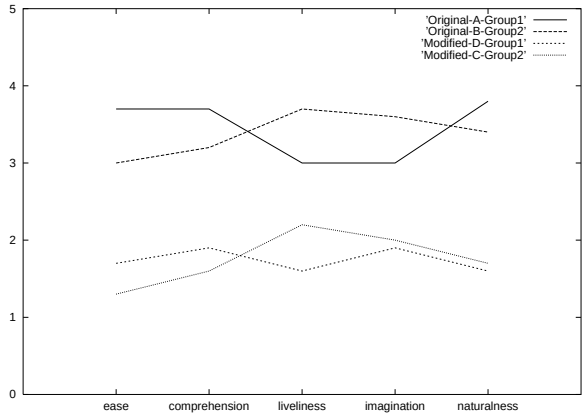

Figure 1: Human sensitivity to pause length

Therefore, if TTS is to be as expressive as human speech, the pause needs be controlled as that of human readings. Our paper aims at revealing the law that governs the pause length and speech content.

3 Hypothesis

When we read out, we need to breathe to obtain the air for our voice. Therefore, the basic reason why we pause when reading out is for breathing. Yet, Sugitou indicates that we control the place to pause so that listeners understand better the content of what is read out. Here are some Sugitou's statements (translated from Japanese):

- When we talk, we unconsciously try to match the timing of breathing and the breaks of content to help listeners understand better.
- Pause is very important time for the listener, too. Listeners revise what is told so far (inside his short term memory) and comprehend during this time.

Sugitou has studied that there are three objectives for human to make pauses; to clarify

the syntactic structure; to clarify the semantic structure; and to emphasize. Here again are her statements:

- Reader does not pause because of lack of air. Rather, he controls to pause not to disturb the word or phrase continuity, not to disturb semantic expressiveness, or to subjectively indicate the semantic break. He tries to match his breathing with grammatical break.
- The places of commas and periods, that are explicit grammatical breaks, are deeply related to where pause is taken.
- In order to emphasize, pause can be taken before or after.

As for the text structures, we first have conventional signs such as periods and newlines. Additionally, we utilize the automatic language analysis tools to segment and extract the sentence structure. As for the emphasis, on the other hand, there are no automatic routine to tell us where in text should be emphasized. However, as Sugitou talks about emphasis differently from structure, emphasis should be something that modify the basic law in between the structure and pause length.

Therefore, we try to find the relationship between the structure of text and pause. From the above Sugitou's statements, we may say that pause is taken between the break of two text elements. Here, the elements can be characters, words, phrases, sentences and chapters. From her reflection on "continuity", probably longer pause is taken between the larger text element. With this hypothesis, however, we could only compare pause of different element sizes. What we want to know also is how pause length varies given two pairs of the same text elements. Therefore, we set our hypothesis to be:

> **When the two text elements are related more tightly, the pause in between is shorter.**

Our hypothesis must be rather natural for all readers. Think of two pairs of sentences A and B that share a subject whereas C and D have different subjects. The reader reads out A and go on to B after a pause to indicate that a sentence has ended, but not too much to show that B is the related sentence. In the case of C and D, on the other hand, the reader wants to stop more in between, to indicate that C and D talk about the different topics.

But how should we define the relationship between two elements? We decided to define them according to each element as follows:

- **Chapters, paragraphs:** cosine value of word frequency vectors
- **Sentences:** subject sharing
- **Phrases:** distance to the word that a both elements modify

Off course, there could be many other possibilities to define relationship between elements. We chose the above not only because they show one aspect of the test relations, but also because they are possible to be processed automatically. Therefore, if we could say that the above hypothesis is true for these relationships, then, we could automatically make the TTS to read out text with more natural pauses, because language analysis could be performed automatically.

4 Data and its Processing

4.1 Data

Among various kinds of speech data, we chose Japanese story, read by a professional story teller in monologue. We selected data that meet the following demands:

- Data is read lively with emotion.
- The reading is affected only by the text, not from any outside reasons.
- Data contains less noise and it has no background music. This constraint is put so that that a pause could be defined as the period without any sound, and also that the data goes thorough the automatic signal processing.

The data we chose are indicated in Table 1. Three monologues are read each by an professional male story tellers.

Next, the electronic text of whole story was obtained and was fed into two text-to-speech software[8][9]. The resulting speech data is also analyzed to be compared with that read by humans.

Table 1: Three monologue data used for analysis

data	data1	data2	data3
title (in Japanese)	Kappa	Bocchan	Kumono-ito
content type	ironical	comical	instructive
length	31min57sec	42min49sec	11min38sec
No.chapters	6	4	3
No.paragraphs	41	22	14
No.sentences	252	431	61
No.phrases (*bunsetsu* of Japanese)	2281	3798	739
No.occurrences words	5016	8379	1644
No.different words	1083	1588	421

4.2 Extraction of Pause

In order to analyze the relationship between the pause and text, we first need to align pause locations with text. We explain how we did this.

The text is analyzed into elements denoted in §3. Chapters, paragraphs, sentences are detected from the original text structures. We used all of them in the analysis. As for phrases, we first used Japanese parser[6] and obtained the sentence structures, because the pause between phrases are related to the sentence structures. We checked all the sentences and extracted the relevant ones, and the ones without any structural ambiguity. From these sentences, we extracted phrases[1] that has pause just afterwards. It left us about one fifth of *bunsetsu* shown in Table 1.

At the same time, the speech data is analyzed using Fourier transformation. Because we chose speech data with least noise, the max power of the Fourier transformation becomes near to 0 when pause is taken. We first extracted all places of max power under a certain threshold locating them as pauses. The result forms a sequence of offsets in the speech data.

Now we have two sequences, sequence of text elements, and sequence of pauses. We aligned these two sequences in two stages. First, sequences are aligned at the sentence level with the aid of dynamic programming. Then, we aligned at the phrases level. Finally, we manually checked and modified all alignments before we went on to the actual analysis.

[1] We call *bunsetsu* as phrase in this paper. *Bunsetsu* is the smallest unit of language in Japanese speech. It is formed of words and several particles, similar to short phrases in English.

5 Analysis Results

5.1 Element Size and Pause Length

Table 2 shows the statistics of pause lengths after each text element for all three data . We see that the average length increase according to the text element size. On the other hand, the deviation does not change up from the paragraph size.

Note that chapter breaks also corresponds to paragraph breaks, sentence breaks and phrase breaks. Because two smaller elements are coupled tighter than the larger ones, this fact suggests that the pause is shorter when two elements area related tighter. Therefore, we may say that our hypothesis holds so far.

We transformed this data into the graphFigure 2. Also, we plot the same lines obtained by the analysis of the same data read by TTS. The horizontal axis shows the number of characters. The vertical axis denotes the pause length. The curve was smoothed so that we could see the trends better. We clearly see the difference of TTS and the human readings. The pause length is the same for sentence, paragraph and chapter for TTS. Also, we found that TTS always have very small variance of pause length.

Table 2: Pause after each text element(Human)

	Avr.Len. (ms)	Deviation.Len.
Chapter	4780	1302
Paragraph	2609	1550
Sentence	1444	642
Phrase	82	206

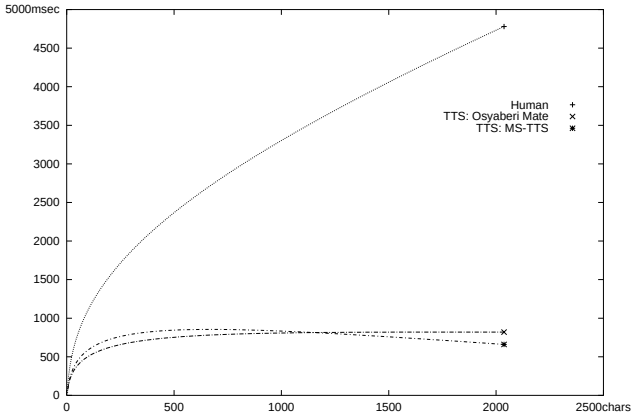

Figure 2: Average and deviation of pause length

5.2 The Relation of Elements and Pause Length

We are now interested in what makes pause length vary. This is in fact exactly the same as discussing our hypothesis. What makes difference in between pairs of text elements with longer pause and shorter pause? In the followings, we show the results of the measures proposed in §3.

Phrase

We measure the phrase relations by the minimum distance between the phrases and the phrase that they modify in common. For example, having the sequence of phrases A B C D E, A modifies D, B modifies C, C modifies D, D modifies E, then the minimum distance is the number of phrases between A and D, that is 3. This in fact is equal to the distance between the first occurring phrase and the phrase it modifies (A on D). This fact, the minimum distance equals to the distance between the first phrase and its modifying phrase, nearly always

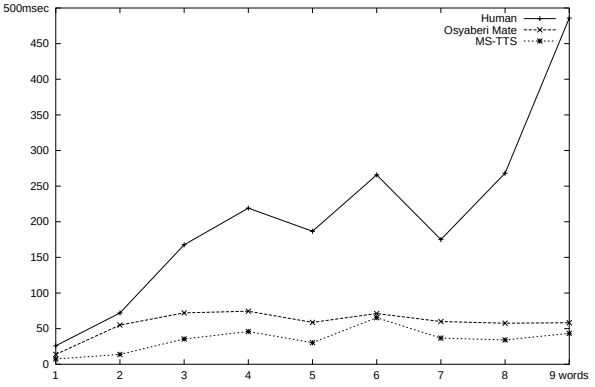

Figure 3: Relation of phrases and pause length

holds in Japanese (therefore, A modifying C and B modifying D rarely occurs).

We obtained pairs of phrases of distance n and obtained the average pause length for each n. This was a hard task from two reasons; first because the frequency of long sentences are very low in the original text; secondly because the percentage of the parser error gets higher for the longer sentences. Therefore, we checked all parse trees by hand and filtered out incorrect ones. These facts left us only a few to 30 examples for $n > 4$. On the other hand, for $n \leq 4$, we obtained at least more than 50 examples.

Figure 3 shows the relationship of phrase distance (horizontal axis) and pause length (vertical axis). Up to 4, we clearly see that human takes longer pause when n increase. However, this trend stops at distance 4 and suddenly turns into a zig zag. We may think of two reasons for this observation. First is that the number of examples was too few to obtain the reasonable average values. This is guessed to be quite true to see the zig zag trend in the figure up from distance of 4. Second reason is that distance of 4 might be the maximum distance for human to be able to read ahead and comprehend the phrase structure while reading out text.

We also plotted the lines of TTS. TTS had the same pause length whatever the distance of the two successive phrases was.

We also examined phrases that had pause longer than average shown in the same figure. Many of them were when one of the phrases was

emphasized, just as Sugitou has indicated.

Sentence

The distance of two sentences are regarded to be small if they contain the same subject. We took top 100 pairs of two sentences with short pauses and judged whether the subjects of sentences are shared or not. These 100 pairs have the average pause of 77.6msecs (whereas the average pause for all sentence breaks was 144.4msecs), therefore clearly shorter pause was taken in between these 100 pair of sentences.

The subjects were shared at 60 % rate. Those sentences with different subjects were the two subjects of conversation scene. We can therefore say that in most cases, the two successive sentences with shorter pause share the context.

The same is done for the pairs with long pauses. The average was 234.5msecs. This time, the subject was shared only at 26 % rate. Therefore, we may say that our hypothesis holds for the sentences.

Analysis on these 26 sentences were made. Although two sentences share the subject, the pause was mostly long because of emphasis. For example, the former sentence draws in a subject and make longer pause to emphasize the subject.

To resume, pause length between sentences is shorter when two sentences are related by context. Especially, the sharing of subject helps to decide the pause length. Also, we found that emphasis on the sentence is expressed by breaking this law, such that longer pause is taken than default if the subject needs to be emphasized.

Paragraph

The distance between two paragraphs is measured by the cosine measure of two word frequency vectors of the paragraphs. Figure 4 shows the plots of pause length (vertical axis) for a cosine measure (horizontal axis) of the pair of two paragraphs. We expected that the plots form a decreasing line from the left top (long pause for small cosine value) to the right bottom (short pause for the large cosine value). Although data3 shows this tendency, data1 and

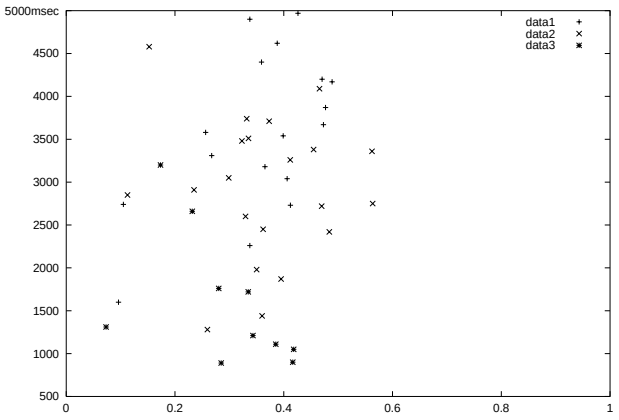

Figure 4: Paragraph relation and pause length

data2 describes that the pause length and cosine value is not correlated. As the cosine value is proved to be the value useful to detect text structure as is shown by [2], this results show that pause length does not correlate with semantic relation of paragraphs.

We could think of a reason for this. According to Sugitou's statement, she mentioned about "short term memory" of listeners. When the human reader is reading out text linearly, a paragraph is too large a structure to be handled in the listeners' short term memory. Therefore, even if the reader could control pause length according to the paragraph continuity, the listener cannot catch such a detailed meaning. Therefore it is useless for the reader to control pause at paragraph level. Rather he just clearly shows the break of paragraph using larger pause than that of sentences.

6 Towards TTS with More Natural Pausing

According to our results, much could be done to make TTS to sound more natural. First, pause length can be controlled to be similar to that of human averages. Secondly, the distance of successive text elements can be measured automatically, and then if the distance is smaller, we may vary the pause length.

We are now currently working on a preprocessing software that automatically puts TTS commands into the text. The preprocessor puts

command with the parameter of probable pause length that is obtained by our analysis.

7 Conclusion

With the objective to make TTS more lively, we analyzed speech data focusing on pause length. With our hypothesis of "when two text elements are related more tightly, the pause in between is shorter", we made experiments on Japanese speech data of story telling. We first saw that longer pause is taken between larger text elements. Then, according to our definition of relationship between text elements such as paragraphs, sentences and phrases, we verified whether the pause length vary according to the tightness of their relations. This was true for the smaller elements up to sentence, not true for the paragraph. Additionally, we also found out that the emphasis is expressed by taking the longer pause than default length, breaking our hypothesis. We also compared the human readings with those of TTS for the same text and found out that TTS have very different naive trends.

There are two important future works. First, we should verify whether our results hold for other speech data. Secondly we should invent a framework to estimate the probable pause length according to the text that comes before and after. Then, we build it into a pre-processor to cause the TTS to sound more natural.

References

[1] S. Pan and J. Hirschberg. Modeling Local Context for Pitch Accent Prediction. ACL, 2000.

[2] M.A. Hearst. Multi-paragraph segmentation of expository text. ACL, 1994.

[3] C. Nass and B. Reeves. Media Equation. Cambridge University Press. 1999.

[4] J. Fry. F_0 Correlates of Topic and Subject in Spontaneous Japanese Speech. ICASSP, 2000.

[5] M. Sugitou. Let's read out aloud(in Japanese). Meiji Shoin. 1996.

[6] S. Kurohashi. The manual of Kyoto Nihongo Parser. 1993. http://www-lab25.kuee.kyoto-u.ac.jp /nl-resource/knp.html

[7] S.Kurohashi. The manual of Juman, the Japanese morphological Analyzer. 1992. http://www-lab25.kuee.kyoto-u.ac.jp /nl-resource/juman.html

[8] Fujitsu. Oshaberi Mate, An Japanese Text-To-Speech Software http://www.fmw.co.jp/s11/oshaberi /oshaberi_s.html

[9] Microsoft TTS Home http://msdn.microsoft.com/workshop /imedia/agent /default.asp