AUTOMATIC ESTIMATION OF ACCENTUAL ATTRIBUTE VALUES OF WORDS TO REALIZE ACCENT SANDHI IN JAPANESE TEXT-TO-SPEECH CONVERSION

Nobuaki MINEMATSU†, Ryuji KITA†, and Keikichi HIROSE‡

†Graduate School of Information Science and Technology, University of Tokyo
‡Graduate School of Frontier Sciences, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 JAPAN

ABSTRACT

Accurate estimation of accentual attribute values of words, which is required to apply rules of Japanese word accent sandhi to prosody generation, is an important factor to realize high-quality text-to-speech (TTS) conversion. The rules were already formulated by Sagisaka *et al.*[1] and are widely used in Japanese TTS converters. Application of these rules, however, requires values of a few accentual attributes of each constituent word. In this paper, these values were estimated through a long series of listening experiments. Here, collection of data of accent types of accentual phrases and estimation of the attribute values from the phrase data were done, where inter-speaker differences of knowledge of accent sandhi were well considered. The rules were further modified to improve the coverage over the obtained data. Evaluation experiments showed the high validity of the estimated values and the modified rules.

1. INTRODUCTION

Several functions, such as grapheme to phoneme conversion and speech waveform generation, need to be implemented to realize a TTS conversion system. Among them, generation of prosodic features from input text is very important and requires a sophisticated process, since no information on prosody is directly given in the text. In the case of Japanese, control of fundamental frequency (henceforth F_0) movement is very crucial to achieve high quality in synthetic speech. In order to realize a good control for a given text, location of an accent nucleus should be adequately estimated for each accentual phrase as well as boundaries of prosodic clauses (breath groups), prosodic phrases, and accentual phrases.

An accentual phrase is often composed of two words or more, typically a content word followed by a function word. Although all the content words (and some function words) have their own accent nucleus positions as their lexical attributes, accent nuclei of accentual phrases shift in many cases due to accent sandhi. These shifts have to be correctly predicted in TTS converters. Accent sandhi rules can be found in an accent dictionary such as [2] but they are in abstract form and not adequate to be used for TTS conversion. Sagisaka *et al.* formulated these rules in a good shape[1], which are widely adopted in Japanese TTS converters. Application of the rules in the system development requires a database of values of several accentual attributes of every word in the vocabulary. Since a public database of the accentual attribute values does not exist and therefore, each development site has had to build the database independently, which is often a time-consuming task.

In the current paradigm of research and development of speech technology, it can be often seen that various tools, databases, and even source codes in some cases are distributed and shared among different sites. Due to this paradigm, especially in speech recognition area, the development efficiency and the system performance have been drastically improved. A Japanese project of "Development of fundamental softwares of anthropomorphic agents" was organized with the aim of developing the agent softwares and distributing them[3]. One of the main goals of the project is to distribute modules and databases required to develop a Japanese TTS converter. The presented work was arranged for this project.

It should be noted that the accentual attributes focused here are NOT accent types of Japanese words, which can be easily obtained by looking up an accent dictionary or asking native speakers of Tokyo dialect. The attribute values to be estimated here are rather difficult even for native speakers to guess only by their introspective considerations of properties of the Japanese language. They have to be estimated experimentally by collecting a large amount of data of accent types of accentual phrases and estimating the values with the phrase data. Recently, there are many cases where researchers should develop a system of foreign tongues. In these cases, construction of databases requiring a deep knowledge of the target language often blocks the efficient development. One of the objectives of this work is to help realizing a common research platform of Japanese TTS converter development by providing a sharable database of the accentual attribute values.

2. WORD ACCENT SANDHI RULES OF JAPANESE

2.1. Word accent of Japanese

Word accent is one of lexical attributes specific to each word and, in Japanese, it is represented by a sequence of binary F_0 levels (H/L) in mora unit. Although it implies 2^N different accent types for N-mora words, the number of accent types for N-mora words of Tokyo dialect is reduced to N+1 due to the following properties.

1. A rapid rising or falling of F_0 has to occur between the first mora and the second one.
2. The number of the rapid falling pattern(s) of F_0 between two consecutive morae is one at most.

Accent type showing a rapid downfall of F_0 just after the n-th mora is called type-n word accent and the n-th mora in this case is called accent nucleus. Figure 1 shows four accent types of 3-mora words and accent nuclei indicated by filled black circles.

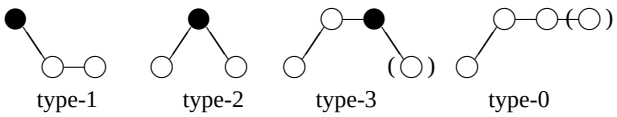

Fig. 1. Accent types observable in 3-mora words

2.2. Word accent sandhi rules of Japanese

When a word is concatenated with another to form an accentual phrase, the resulting position of the accent nucleus of the phrase is often different from any of original nuclei of the constituent words. The word accent sandhi can be categorized into three types;

1. **Shift** of the accent nucleus
   ア┐カ + エンピツ → アカエ┐ンピツ
   red pencil
2. **Generation** of the accent nucleus
   ケイタイ + デンワ → ケイタイデ┐ンワ
   portable telephone
3. **Deletion** of the accent nucleus
   ケ┐イザイ + テキ → ケイザイテキ
   economy (suffix) economical

Development of Japanese TTS converters requires rules of the word accent sandhi, which were well formulated by Sagisaka *et al.*[1]. The following sections briefly describe the rules, which are composed of three set of rules and several control rules over the three. Three accentual attributes of concatenation manner (CM), nucleus position (NP), and concatenation type (CT) are defined for the rules. Values of the attributes of words are estimated in this study.

2.2.1. Concatenation of a content word and a function word

Suppose that concatenation of a content word of N_1 morae and type-M_1 accent and a function word (an auxiliary verb or a particle) of N_2 morae and NP being $\widetilde{M}_2$ produces an accentual phrase of N_c morae and type-M_c accent. NP is an attribute indicating accent nucleus location in the produced accentual phrase. If the accent nucleus is located as the last mora of the first word, NP is zero. If the first mora of the second word is the accent nucleus, NP is one. It should be noted that NP can have negative values.

If every word which can appear as the second word has its own value of NP, CM is not needed. This is because location of the accent nucleus can be determined only by NP. However, in some cases, the accent nucleus of the first word remains after the concatenation. In these cases, the nucleus position of the phrase cannot be predicted only by accentual attributes of the second function word. To sum up, it can be said that the accent nucleus position of an accentual phrase composed by a content word and a function word is determined by length and accent type of the first word and CM and NP of the second word. Table 1-(a) shows these word accent sandhi rules. As shown in the table, all of the four factors above are not always required to determine the nucleus location.

2.2.2. Concatenation of two content words

Word accent sandhi observed when concatenating two content words can be characterized by adequately setting CM and NP values of the second *content* word. It means that these values have to be estimated for every content word. But when the second word is a verb or an adjective, the accent nucleus of the resulting phrase is always found as the last mora but one in the phrase (M_c=N_1+N_2−1). This property of Japanese requires the estimation only when the second word is a noun. In this case, unlike function words described in the previous section, since CM value of the second noun word is always F4 or F5, NP value has only to be estimated for each concatenated noun. Table 1-(b) shows the word accent sandhi rules in concatenating a content word and a noun. Although concatenation types (CT) are newly defined in the table, they are functionally the same as NP. C1 to C4 correspond to NP values of M_2, 1, 0, $-N_1$ respectively. Since NP values of nouns of three morae or longer can be deterministically obtained by their length and accent types, only the nouns of two morae or shorter were examined.

Table 1. Word accent sandhi rules of Japanese

word of N_1 morae and type-M_1 accent +
word of N_2 morae and nucleus position (NP) being $\widetilde{M}_2$
→ accentual phrase of N_c morae and type-M_c accent

(a) concatenation of a content word and a function word

concatenation manner	M_c	
	$M_1 = 0$	$M_1 \neq 0$
(F1) 従属型 *	M_1	
(F2) 不完全支配型 *	$N_1 + \widetilde{M}_2$	M_1
(F3) 融合型 *	M_1	$N_1 + \widetilde{M}_2$
(F4) 支配型 *	$N_1 + \widetilde{M}_2$	
(F5) 平板化型 *	0	

(b) concatenation of two content words

concatenation type	conditions of the second word	M_c
(C1) 保存型 *	$N_2 \geq 2, M_2 \neq 0, N_2{}^{\dagger}$	$N_1 + M_2$
(C2) 生起型 *	$N_2 \geq 2, M_2 = 0, N_2{}^{\dagger}$	$N_1 + 1$
(C3) 標準型 *	$N_2 \leq 2$	N_1
(C4) 平板型 *	$N_2 \leq 2$	0

(c) concatenation of a prefix and a content word

concatenation type	M_c	
	$M_2 = 0, N_2{}^{\dagger}$	$M_2 \neq 0, N_2{}^{\dagger}$
(P1) 一体化型 *	0	$N_1 + M_2$
(P2) 自立語結合型 *	$N_1 + 1$	$N_1 + M_2$
(P3) 分離型 *	M_1	M_1 (and $N_1 + M_2$)
(P4) 混合型 *	$N_1 + 1$ (or)M_1	M_1 (and/or) $N_1 + M_2$

† : If the final syllable of the second word is comprised of two morae, N_2 should be decremented by one.

∗ : In Sagisaka's original paper in Japanese, as shown here, each value of CM and CT has a meaningful name, not a label. Due to limited space, however, these values are referred to by labels of Fx, Cx, and Px, hereafter.

2.2.3. Concatenation of a prefix and a content word

To make an accentual phrase by attaching a suffix to a content word, the rules in Section 2.2.2 can be basically applied as they are. For a phrase composed by a prefix and a content word, new rules should be prepared, which are shown in Table 1-(c). It should be noted that, for P3 and P4, semantic analysis is sometimes required to adequately locate the accent nucleus.

2.2.4. Control rules over the three sets of concatenation rules

To make an accentual phrase from two words, the above three sets of rules are selectively used according to part-of-speech of the two words. However, several additional control rules have to be referred to in their actual application to TTS converter development.

- To concatenate three words or more into one accentual phrase, the above rules are applied recursively from left to right.
- If the resulting accent nucleus is found at a syllable with a Japanese *tokushuhaku* phoneme, such as a long vowel, a double consonant, or a syllabic nasal, the nucleus is shifted leftward by one mora.
- If the resulting accent nucleus is located at a mora with an unvoiced vowel sound, the nucleus is shifted leftward by one mora.
- In the case of *ichi-dan* verbs in the form of negative or adverbial root, the nucleus is shifted leftward by one mora after applying the above rules.

3. PRELIMINARY EXPERIMENTS TO ESTIMATE VALUES OF THE ACCENTUAL ATTRIBUTES

As told in Section 1, values of the accentual attributes shown in Table 1 are very difficult even for native speakers to tell by their introspective considerations of properties of the Japanese language. By carefully and intensively looking at the rules, however, a deterministic procedure can be derived to estimate values of the attributes. Figure 2 shows the estimation procedure when concatenating a content word and a function word. In the figure, Fx/N means that CM is Fx and NP is N. Ten university students were asked to follow this procedure to estimate the attribute values of about 1,400 words. Before the estimation experiment, a listening test was done to check whether the subjects could indicate accent types correctly from an utterance of a *meaningless* sequence of morae. This is because it is known that some people cannot indicate change of tone correctly though they can perceive it. Rate of subject i's correctly indicating the accent type is denoted as ω_i and it is used as reliability measure of the subject. After the experiment, it was often found that different subjects estimated different values for a word. Then, using ω_i, reliability r of estimated value λ for each target word was introduced, which was defined as

$$r(\lambda) = \frac{\sum_{\text{subject } i\text{'s response} = \lambda} \omega_i}{R}$$

where R is summation of ω_i for all the subjects irrespective of their estimated values. Analysis was done for the estimated value with the highest reliability of each function word. Figure 2 shows the ratios of function words with their highest reliability being larger than θ (preliminary estimation). The figure also shows the ratios obtained by the new estimation method, which will be described later (proposed estimation). It has to be admitted that subjects' responses have large variance and the estimated values have very low reliability, which is due to the following reasons.

- Procedures required good knowledge of Japanese grammar and phonology. After the procedures in Table 2 were carried out, the control rules in Section 2.2.4 had to be considered. In this step, differences in the subjects' knowledge on Japanese grammar and phonology is supposed to have enlarged variance among the subjects' responses.
- Estimation of the attribute values of a given target word requires a series of procedures to be done. So, the number of different content words used for the estimation was set to small, which caused low reliability.
- To reduce task of the subjects, only a few content words were selected and given to the subjects for all the target non-content words. In some cases, concatenation of one of the selected content words and a target word produced a meaningless phrase, which confused the subjects.

Table 2. Procedures of determining accentual attribute values of a given auxiliary verb

1. Concatenate verb "あるく"(with a nucleus) and the auxiliary verb.
   If accent type of the phrase is 0 (no nucleus), then the value is F5.
   else if it is 2, then go to step 2.
   else if it is 3 (mora-based length of "あるく")+N, go to step 3
2. Concatenate verb "わらう"(no nucleus) and the auxiliary verb.
   If accent type of the phrase is 0 (no nucleus), then the value is F1.
   else if it is 3+N, then the value is F2/N.
3. Concatenate verb "わらう"(no nucleus) and the auxiliary verb.
   If accent type of the phrase is 0 (no nucleus), then the value is F3/N.
   else if it is 3+N, then the value is F4/N.

あるく is walk and わらう is laugh. You can use any other verbs instead of these two if they consist a pair; one is with a nucleus and the other is without it when uttered in isolation.

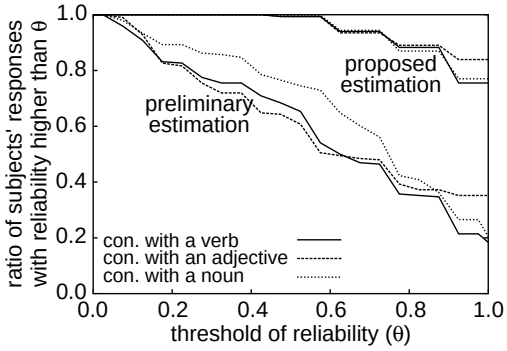

Fig. 2. Ratio of subject's responses with the highest reliability being larger than θ

4. IMPROVED ESTIMATION OF VALUES OF THE ACCENTUAL ATTRIBUTES

Considering the defects of the previous experiment, another experiment was carried out, which was designed so that the subjects' task was easy, results of ample estimation trials were available, and the estimated values should be highly reliable.

4.1. Improved experimental setup

4.1.1. Selection of the subjects and their new task

To realize highly reliable estimation, 7 whose ω_i scores were larger than 0.9 were selected. Their task was simply to answer the accent type of an accentual phrase given as text. Since it will be easier for the subjects to answer the accent type than to follow the procedures in Table 2, more reliable and consistent results will be available. Estimation of the attribute values were carried out by an algorithm proposed in Section 4.2 which considered quite well different accent types estimated by different subjects.

4.1.2. Preparation of accentual phrases

Accentual attribute values were estimated for approximately 1,600 words. Here, about 200 3-mora words were newly added, since the previous experiment indicated their necessity. Accentual phrases including these words were selected from five years' newspaper text database lest meaningless phrases would be presented to the subjects in the experiment.

For concatenation of a content word and a function word, verbs, adjectives, and nouns were considered as content words. For each function word, four accentual phrases comprised of a verb and the function word were selected; two having verbs with an accent nucleus when uttered in isolation and the other two having verbs without it. As for adjectives and nouns, four examples were also selected for each function word in the same way as they were done for verbs. In the case of prefixes, twelve phrases were prepared for each prefix, six having content words with an accent nucleus (two verbs, two adjectives, and two nouns) and the other six having content words without it. Unlike function words and prefixes, for each suffix, only six phrases (two verbs, two adjectives, and two nouns) were prepared irrespective whether the content word had an accent nucleus or not. Finally, the total number of different accentual phrases prepared for the experiment was about 4,200. The total time for listening to them and identifying their accent types was 30 hours per subject, 210 hours all together.

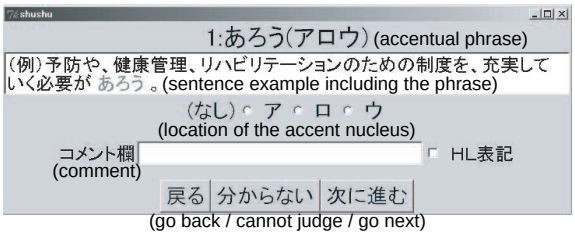

Fig. 3. GUI interface of a tool built for the data collection

4.1.3. Confirmation through listening tests

Answering an accent type of a given accentual phrase is possible only by text input. However, what we need is the data of accent types which *sound* natural. Therefore, in the experiment, after the text-based answering, confirmation task was introduced, where synthetic speech was presented with the accent nucleus located as indicated by the subject. Figure 3 shows GUI of a tool built for the data collection. An accentual phrase was shown with a sentence example including it. The subjects were asked to click an accent nucleus position which they thought was correct. After that, a synthetic speech sample was generated with the selected accent type. If the sample sounded natural, they clicked OK to go to the next phrase. Otherwise, they had to go back to the answering step.

4.2. Algorithm to estimate the accentual attribute values from the obtained data

This section describes an algorithm to estimate the accentual attribute values from the obtained data of accent types of accentual phrases. Here, by taking まで (/made/) as an example of a particle, the estimation algorithm for verb concatenation is completely described. In the case of the other concatenations as well as the other non-content words such as auxiliary verbs, prefixes, and suffixes, the proposed algorithm can be applied only with a minor modification. Table 2 tells us that the word accent sandhi manners in the case of concatenating a content word and a function word can be divided into two cases. One is that the content word is with an accent nucleus uttered in isolation and the other is that the content word is without it. Considering this property, four accentual phrases, two having verbs with an accent nucleus and the other two having verbs without it, were prepared from the text database. Table 3 shows actual data obtained for まで. Using the data, the following steps were automatically carried out for the estimation.

1. Out of the four verb phrases, four phrase pairs are formed, each pair of which has a phrase whose verb is with a nucleus and the other whose verb is not when uttered in isolation. Pick up a phrase pair. For the phrase pair, response pairs are defined. A response pair includes a single subject's responses for both of the two responses. We have 49 different response pairs since we have 7 subjects.

2. For each of the four (verb) phrase pairs, do the following.

   For each of the 49 response pairs, by using the procedures of Table 2, obtain the accentual attribute value of the selected phrase pair, which can be done deterministically. When the phrase pair is (でるまで, するまで) and the response pair is (s1,s1), the value of まで is determined to be F2/0.

   The phrase pair often results in having several different accentual values. In these cases, by considering how many response pairs indicate each of the candidate values, a validity score is assigned to each value. This score is ranged from 1/49 to 49/49.

Table 3. Subject's responses for verb phases including まで

presented phrases		subjects' responses (accent types)						
		s1	s2	s3	s4	s5	s6	s7
$M_1 \neq 0$	でるまで (M_1=1)	1	1	1	1	1	1	1
	いたるまで (M_1=2)	2	2	2	2	2	2	2
M_1=0	しぬまで	2	3	3	3	2	3	3
	するまで	2	3	3	3	3	2	3

でる, いたる, しぬ, and する are verbs.

3. Different phrase pairs often give us different sets of accentual values with the validity scores. Out of all the different accentual values, the value with the highest validity score is selected.

For some responses, the attribute values were not obtained with the rules in Table 1. For these responses, new attributes or values were introduced. Due to limited space, the new rules are not described here. Interested readers should refer to literature[4].

4.3. Evaluation of the estimated values and conclusions

Two evaluation experiments were carried out with respect to the estimated values. Firstly, we calculated the reliability of the subjects' responses by using a similar method to the $r(\lambda)$ calculation. In the experiment of the improved design, the subjects' responses were not accentual values but accent types. Figure 2 also shows the ratio of responses with their highest reliability being larger than θ. Clearly shown in the figure, very high reliability is realized and this indicates the validity of the experimental design discussed in Section 4.1. Next, accuracy of predicting accent types of accentual phrases with the estimated attribute values were examined with two databases, ATR sentence database and Kyoto university text database. Out of the two databases, 780 two-word phrases each of which included an auxiliary verb, 1,583 phrases each including a particle, 470 phrases each including a prefix, and 2,177 phrases each including a suffix were selected. Using the estimated values, accent types of the phrases were predicted and the predicted types were evaluated by three native speakers of Tokyo dialect. Accuracy rates were 96.2 %, 91.5 %, 93.6 %, and 92.3 % for each case. Although the new rules were rarely used for auxiliary verbs, particles, and suffixes, 40 % of the phrases with prefixes required the new rules. Error analysis indicated that 90 % of the prediction errors were found in phrases which appear only in classical style of writing. Since the rules in Table 1 don't consider classical dialects, these errors are inevitable. If we use the rules in Table 1 and the new rules with the estimated values to convert contemporary Japanese sentences to speech, we can conclude that the obtained database is highly useful, effective, and reliable.

5. REFERENCES

[1] Y. Sagisaka and H. Sato, "Accentuation rules for Japanese text-to-speech conversion," Review of the Electrical Communication Laboratories, vol.32, no.2, pp.188-199 (1984).

[2] "Word accent dictionary of Japanese pronunciation," published by NHK (Nippon Hoso Kyokai) (1998).

[3] S. Kawamoto *et al.*, "Design of software toolkit for anthropomorphic spoken dialogue agent software with customization-oriented features," Journal of Information Processing Society of Japan, vol.43, no.7 (2002, to appear).

[4] R. Kita, N. Minematsu, and K. Hirose, "Development of rules of word accent sandhi and their improvement for Japanese TTS systems," Technical report of Institute of Electronics, Information and Communication Engineers, SP2002-26 (2002).