

ON THE LEVELS OF ACCENTUATION IN SPOKEN JAPANESE

Hiroya Fujisaki, Sumio Ohno and Osamu Tomita

Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda, 278 Japan

ABSTRACT

Accentuation serves to express both the discrete information concerning the accent type of a prosodic word and the continuous information concerning its prominence. This paper examines the latter aspect of accentuation using recorded radio news read by announcers. The amplitude of the accent command was extracted from an F_0 contour and used as an index for the level of accentuation. Statistical analysis of the accent command amplitude confirmed the difference between accented and unaccented types. Further analysis of the relationship between amplitudes of two adjoining accent commands also revealed a marked difference in characteristics of these two types.

1. INTRODUCTION

Prosody can be defined as the systematic organization of various linguistic units into an utterance or a coherent sequence of utterances in the process of speech production by the speaker, and it serves to facilitate comprehension of the spoken messages by the listener. As far as the spoken Japanese is concerned, the organization is accomplished basically by three means: accentuation, phrasing, and pausing.

Accentuation in the spoken Japanese is defined as the assignment of one of the allowed binary (high/low) patterns of subjective pitch to a sequence of morae in an utterance. It serves for the listener to decide whether or not a given sequence of morae can be a candidate for a lexical item or a sequence of lexical items that are syntactically closely tied together such as a noun and a particle. Thus accentuation evidently expedites lexical access. The scope of accentuation, i.e., the constituent(s) to which a word accent type is assigned, is defined as a *prosodic word*.

Phrasing is defined as the assignment of a slowly declining pattern of subjective pitch, often accompanied by a reduction in the local speech rate at the end and occasionally but not necessarily accompanied by a pause, to the whole or a part of an utterance that constitutes a syntactically coherent unit such as an immediate constituent with recursive left branching. Obviously, phrasing facilitates parsing. The scope of phrasing is defined as a *prosodic phrase*.

Pausing literally means to insert a silent interval within an utterance or between utterances. A pause generally marks the end of a major syntactic constituent such as a clause within a sentence as well as the end of a sentence within a discourse. The length of pause is controlled to indicate the size of the constituent. Pausing provides the listener with the time necessary for higher level syntactic and

semantic processing, while serving the need for breathing on the part of the speaker.

All the three means of marking the prosodic units produce objective and quantifiable consequences. The consequence of pausing can obviously be quantified by the duration of a pause except for the final constituent. Consequences of accentuation and phrasing can be quantified by referring to the underlying commands that are extracted from the contour of the voice fundamental frequency (henceforth the F_0 contour). Thus the consequence of accentuation can be expressed by the timing of the onset and offset of the accent command for a constituent, while that of phrasing can be expressed by the timing of occurrence of the phrase command.

In addition to the discrete type of information expressed by their presence/absence and timing, these commands possess another attribute, viz., the amplitude/magnitude, that can convey further information. In the case of the accent command, the timing of its onset and offset relative to the segmental timing is sufficient for expressing the categorical information as to the accent type of the constituent in question, while its amplitude, corresponding to the degree of accentuation, can be used for other types of information.

The present paper addresses the question on the types of information expressed by the amplitude of the accent command, based on the analysis of speech of radio news. Characteristics of the F_0 contours are extracted by the method of Analysis-by-Synthesis using a model for the generation process, and the results are analyzed and interpreted in relation to the underlying information.

2. WORD ACCENT TYPES IN THE COMMON JAPANESE

All the types of the word accent of the common Japanese are characterized by the existence of a transition in the subjective pitch, either upward or downward, at the end of the initial mora of a word, and by the fact that no more than one downward transition is allowed within a word. Thus there exist $(n + 1)$ different accent types in words consisting of n morae. The patterns of subjective pitch, assuming only binary values, are shown schematically in Table 1, which lists all the possible accent types for words that consist of up to 5 morae, along with a sample word for each type [1]. A filled circle in the table represents the subjective pitch of a particular mora, and an empty circle shows that of a particle whose existence serves, in some instances, to discriminate two different types that are otherwise indistinguishable. The abbreviated notation (n, i) is used for the accent type of an n -mora word with a downward transition of pitch at the end of the i th mora. These are often referred to as

Table 1: Patterns of subjective pitch of all the types of word accent in the common Japanese for words that contain up to five morae.

		NUMBER OF MORAE IN A WORD				
		1	2	3	4	5
TYPES OF WORD ACCENT	0	i	o - i	a - o - i	o - i - o - i	a-wa-re-mo-no
	1	o	a - o	a-wa-re	o - i - o - i	a-me-a-ra-re
	2		i - e	a - o - i	i - e - i - e	a - o - i - u - o
	3			u - re - i	a - o - a - o	a - ma - ga - e - ru
	4				o - to - o - to	mi - a - ya - ma - ru
	5					o - sho - o - ga - tsu

the ‘accented’ types. Type $(n, 1)$ is also called the ‘high head’ type (to be abbreviated by HH), while other (n, i) types are collectively called the ‘rise-fall’ types (to be abbreviated by RF). Type $(n, 0)$ is meant to represent the accent type which has no downward pitch transition. This type is often referred to as the ‘unaccented’ type or the ‘flat’ type (to be abbreviated by FL).

Although the above-mentioned classification of accentuation patterns have been developed for classifying prosodic types of isolated words, it can be extended to apply to a string of words which, in connected speech, behaves just like a word as far as accentuation is concerned, hence the name *prosodic word* is adopted.

It is to be noted that the above-mentioned classification is based solely on binary representation of subjective pitch of morae in isolated words, and tells nothing about the quantitative aspects nor about the interaction of accentuation between two or more words in connected speech. Although previous studies by Fujisaki and his coworkers [2] have investigated the influences of the accent type, syntactic structure and focus on the realization of word accent in sequences of two and three prosodic words, the results have been stated only qualitatively. It is the purpose of the present paper to discuss some of these aspects in quantitative terms.

3. THE SPEECH MATERIAL AND THE METHOD OF ANALYSIS

3.1. Speech Material

The speech material for the present study consists of recordings of FM radio news uttered by two male announcers. The text consists of a total of 84 sentences on 18 topics. Because of the particular style common to radio news, the sentences are mostly compound sentences, such as those consisting of several simple sentences connected by conjunctions, and those with multiple embeddings. The total duration of speech is 22 minutes 36 seconds, and the average speech rate is 8.68 mora/sec for one announcer (announcer A) and is 8.37 mora/sec for the other (announcer B).

3.2. Analysis Procedure

The broadcast news were first recorded on a digital audio tape recorder and played back for the purpose of transcription. The speech material was also digitized at 10 kHz with 16 bit precision for further analysis. The fundamental frequencies were extracted at every 10 ms by a modified autocorrelation analysis of the LPC residual. The F_0 contours were further analyzed by the method of Analysis-by-Synthesis using a quantitative model for the process of F_0 contour generation [3].

Figure 1 shows the configuration of the model that have been proved to be valid for F_0 contours of several languages including Japanese, English, German and Spanish. The phrase commands are assumed to be impulses applied to the phrase control mechanism to generate the phrase components, while the accent commands are assumed to be positive stepwise functions applied to the accent control mechanism to generate the accent components. Both mechanisms are assumed to be critically damped second-order linear systems, and the sum of their outputs, i.e., the phrase components and accent components, is superposed on a baseline value $\ln Fb$ to form an F_0 contour on the logarithmic scale, as given by the following equation:

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^I A p_i G p(t - T_{0i}) + \sum_{j=1}^J A a_j [G a(t - T_{1j}) - G a(t - T_{2j})], \quad (1)$$

$$G p(t) \begin{cases} = \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \quad (2)$$

$$G a(t) \begin{cases} = \min[1 - (1 + \beta t) \exp(-\beta t), \gamma], & \text{for } t \geq 0, \\ = 0, & \text{for } t < 0, \end{cases} \quad (3)$$

where $G p(t)$ represents the impulse response function of the phrase control mechanism and $G a(t)$ represents the step response function of the accent control mechanism.

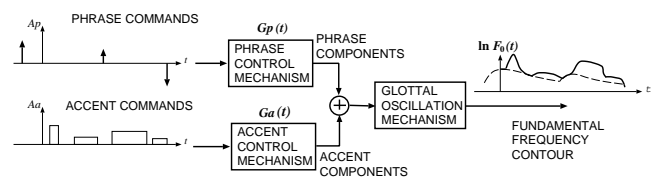


Figure 1: A quantitative model of the process of F_0 contour generation for Japanese, English, German and Spanish.

3.3. Example of Analysis Results

Figure 2 illustrates a typical result of analysis of an utterance by announcer A. Each panel shows, from top to bottom, the speech waveform, the observed F_0 contour as a sequence of + symbols, the best approximation generated by the model as a curve in a solid line, the estimated phrase components as a curve in a dashed line, the estimated baseline value Fb in a dotted horizontal line, the estimated phrase commands as impulses, and the estimated accent commands as stepwise functions. The difference between the solid line and the dashed line corresponds to the estimated accent components. The

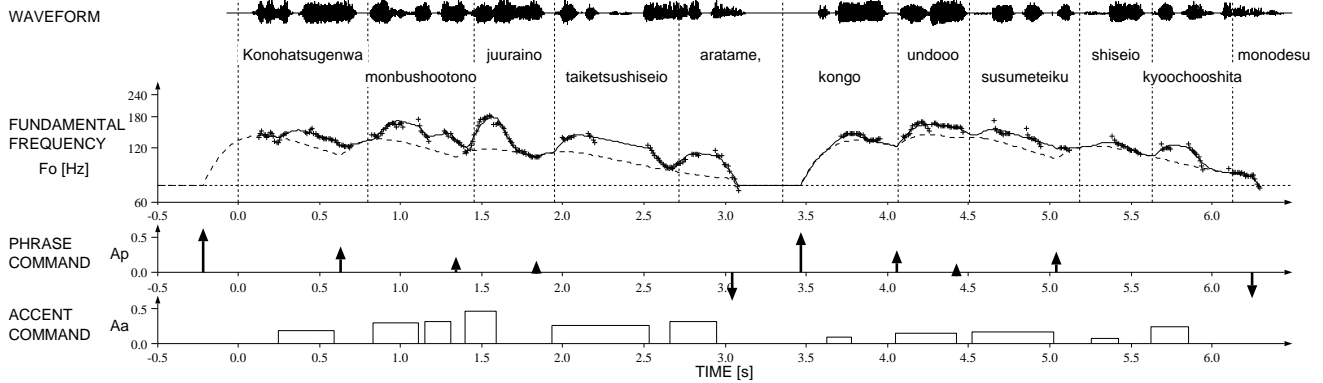


Figure 2: An example of F_0 contour analysis. The utterance is “Konohatsugenwa monbushootono juuraino taiketsushiseio aratame, kongo undooo susumeteiku shiseio kyoochooshita monodesu.” (This statement showed that he intended to change the hostile attitude (of the teacher’s union) of the past against the Ministry of Education, and emphasized the attitude in directing future movements.)

vertical dotted lines indicate boundaries between prosodic words obtained by visual inspection of the waveform and the spectrogram.

As shown by this example, the model is capable of producing very close approximations to the measured F_0 contours, and the commands and resulting components clearly indicate the timing and the magnitude of both phrasing and accentuation, including the complete lack of accentuation on certain words toward the end of an utterance.

4. STATISTICAL ANALYSIS OF ACCENT COMMAND AMPLITUDE

All the utterances have been analyzed to extract, along with other parameters, the amplitude and timing of the accent command for each prosodic word. The influences of various factors such as the accent type, the position within a prosodic phrase, and the accent type of the preceding prosodic word, etc., have been analyzed statistically. Since there exist certain individual differences in the values obtained from the two speakers, the results should not be pooled together but should be processed individually. Because of space limitations, we will show here only the results obtained from utterances of announcer A.

4.1. Influence of Accent Type

The mean μ and the variance σ^2 of the accent command amplitude were calculated for each of the three categories (HH, RF, and FL) of accent types under the three conditions:

- for all accent commands,
- for the initial accent commands in prosodic phrases containing two or more prosodic words,
- for the second accent commands in prosodic phrases containing two or more prosodic words.

These values are shown in Table 2, and the corresponding normal probability density functions are shown in Fig. 3 (a)~(c). These results indicate that the two categories HH and RF are very close to each other but are distinct from the category FL in all three conditions. They also show that the distance between HH/RF and FL are smaller for the condition (c) than for the condition (b), which

is due to reduction of the means for HH/RF and the increase of the mean for FL. Table 3 shows the test results of significance of difference between these three categories, both for the mean and the variance.

4.2. Correlation Between Adjoining Commands

The interaction between amplitudes of two and three adjoining accent commands under all combinations of accent type, syntactic structure, and focus placement has been studied and reported by

Table 2: Mean (μ) and variance (σ^2) of the accent command amplitudes for the three accent type categories HH, RF and FL in the three conditions (a), (b) and (c).

Condition	Mean/var.	Accent type category		
		HH	RF	FL
(a)	μ	0.353	0.343	0.251
	σ^2	0.011	0.016	0.010
(b)	μ	0.388	0.394	0.229
	σ^2	0.013	0.011	0.005
(c)	μ	0.335	0.336	0.289
	σ^2	0.014	0.017	0.010

Table 3: Tests of significance of difference in the three conditions (a), (b) and (c).

Condition	Mean/var.	HH vs. RF	HH vs. FL	RF vs. FL
(a)	μ	-	+++	+++
	σ^2	++	-	+++
(b)	μ	-	+++	+++
	σ^2	-	++	++
(c)	μ	-	-	++
	σ^2	-	-	+

+++ : significant at 1% level + : significant at 10% level
 ++ : significant at 5% level - : not significant

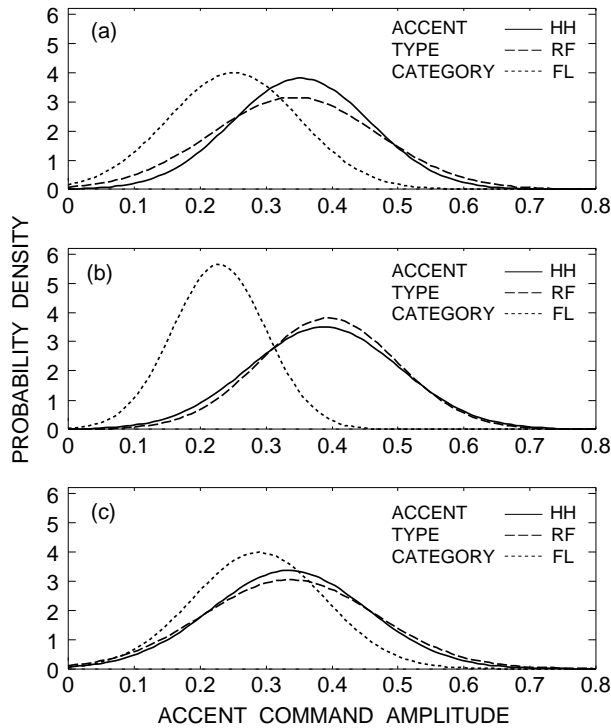


Figure 3: Probability density functions of the accent command amplitude, (a) for all accent commands, (b) for the initial accent commands in prosodic phrases containing two or more prosodic words, (c) for the second accent commands in prosodic phrases containing two or more prosodic words.

Fujisaki and others [2] using utterances in a strictly controlled discourse context. The study indicated that focus placement produces marked and discrete changes in the command amplitude. In order to see if similar changes occur in real discourse, the amplitudes of the initial and the second accent commands (to be denoted by Aa_1 and Aa_2 , respectively), occurring within the same prosodic phrase are plotted in Fig. 4 (a) and (b). Panel (a) shows the case where the initial prosodic word is of the ‘accented’ type (i.e., the HH and RF types), while panel (b) shows the case where the initial prosodic word is of the ‘unaccented’ type (i.e., the FL type). Although it is clear from Fig. 3(b) and 3(c) that, on the average, Aa_2 is smaller than Aa_1 , Fig. 4(a) shows rather low correlation ($r = +0.146$) between Aa_1 and Aa_2 , and indicates no bimodal distribution. When the initial constituent is of the ‘accented’ type, therefore, the current data indicate that both the level of accentuation and the amount of ‘downstep’ (i.e., the decrease in the amount of accentuation from the initial constituent to the second constituent) have broad and continuous distributions. Figure 4(b) also shows low correlation ($r = +0.048$) between Aa_1 and Aa_2 , but indicates a bimodal distribution for Aa_2 . When the initial constituent is of the ‘unaccented’ type, therefore, the current data suggest the existence of a nearly constant value of ‘upstep’ (i.e., the increase in the amount of accentuation). Closer examination of the data reveals that this ‘upstep’ occurs when the second constituent is of the ‘accented’ type. The

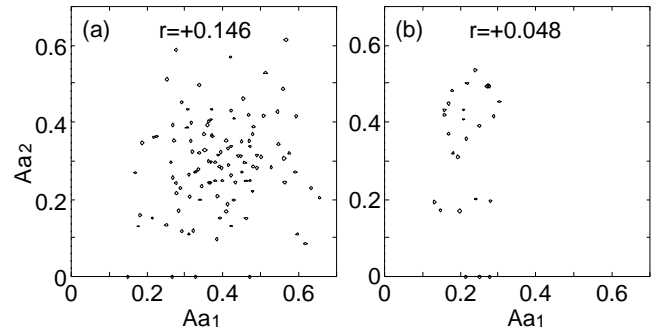


Figure 4: Relationship between accent command amplitudes of the initial and second prosodic words within a prosodic phrase: (a) initial prosodic word: “accented type” (b) initial prosodic word: “unaccented type.”

origin of the apparent difference in the tendency indicated by the two panels in Fig. 4 requires further investigation.

5. SUMMARY AND CONCLUSION

Quantitative aspects of accentuation in speech of the common Japanese has been investigated using recorded radio news. The F_0 contours have been analyzed using a model of the generation process, and the extracted amplitude of the accent command has been adopted as an index for the level of accentuation. Statistical analysis of the distribution of accent command amplitude has confirmed the difference between the ‘unaccented’ type and ‘accented’ types both in the mean and in the variance in most cases. Analysis of the relationship between amplitudes of the initial and second accent commands within a prosodic phrase has also revealed a marked difference in the characteristics of ‘accented’ and ‘unaccented’ types when they are the initial constituent of the prosodic phrase. Further work is under way to investigate the relationship between the objectively measured level of accentuation and the perception of prominence.

6. REFERENCES

1. Fujisaki, H. and Nagashima, S., “A model for the synthesis of pitch contours of connected speech,” *Annual Report of the Engineering Research Institute, University of Tokyo*, 28: 53–60, 1969.
2. Fujisaki, H., Hirose, K., Takahashi, N. and Yoko’o, M., “Realization of accent components in connected speech,” *Transactions of the Committee on Speech Research, Acoust. Soc. Jpn.*, S84: 279–286, 1984.
3. Fujisaki, H. and Hirose, K., “Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation,” *Preprints of Papers, Working Group on Intonation, The XIIIth Int’l Congress of Linguists, Tokyo*, 57–70, 1982.