TR-I-0051

On sentence level factors gaverning segmental duration in Japanese

日本語音韻継続長における文発声固有の性質について

Kazuya Takeda, Yoshinori Sagisaka and Hisao Kuwabara
武田 一哉, 匂坂 芳典, 桑原 尚夫

1988.11

Abstract

In this paper, the durational characteristics of Japanese are statistically analyzed aiming at establishing a fine duration setting rule. First, statistic duration control model is formulated through factor analysis on a large scale isolated word database(5,240 words). Second, using 282 continuous sentence speech, each segmental duration, calculated by this model, is compared with taht of measured one. The error analysis clarified the existence of the following sentence level duration control: 1) Pre-pausal lengthening is greater than simple word final lengthening, 2) Shorter durations are found at sentence final position in declarative sentences. By applying those sentence level controls to the duration setting rule for sentence speech, it was turned out that the prediction error can be reduced to the same amount of errors in isolated speech.

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

Contents

1 Introduction

On the duration characteristics, many research works have been reported in some languages [1-13]. Roughly speaking, those works aim at clarifying inherent length of speech segments and variations of segmental durations in various context. For Japanese speech, Hiki et. al. investigated inherent duration and compressibility of various speech segments and suggested syllable-timed characteristics of Japanese [3,4]. Sato et. al. analyzed time alignment of segments from the stand point of both production and perception, and concluded that the interval between vowel onset of syllables are isochronous [5]. Using those properties of Japanese segmental durations, Sagisaka et. al. and Higuchi et. al. proposed a duration setting rule for speech synthesis and confirmed their effectiveness through experiments [6,7].

Through these studies, such control factors for Japanese segmental duration as those in Table 1 have been proposed. The dominant control factor, neighboring C-V and V-C duration compensation can be considered as the acoustic manifestation of mora-timing. Moreover, word-final lengthening and shortening in inverse proportion to the number of moras in a breath group can be interpreted as the results of the breath group constituting a timing domain.

In natural sentence utterances, however, segmental durations fluctuate more widely than in isolated speech. To establish more natural duration setting rules for speech synthesis, not only word-level factors but also sentence-level factors must be taken into account. Recently, some durational phenomena associated with linguistic environments have been reported for English[10,11,12,13]. These results, such as pre-pausal, phrase-final lengthening and paragraph level controls, suggest the existence of some *supra-word level* duration controlling mechanisms. For Japanese, however, there is almost no statistical data on sentence-level duration controls. This study is aiming at clarifying such sentence-level control mechanisms and proposes a new duration setting rule of Japanese segmental durations.

In section 2, a duration setting rule is proposed based on analysis of a large scale isolated word speech database. As for control parameters for the rule, conventionally used factors such as, type of phoneme, location in a breath group, neighboring phonemes, mora number in a breath group and geminate environment, are used. By using a factor analysis method, the resulting model describes the relationship between duration variations and environmental factors using a simple linear additive equation. Furthermore, an experiment using a 216 word database shows higher accuracy of the model than conventional models.

In section 3, the existence of sentential duration controls is clarified. In this

section, the word-level duration setting rule proposed in section 2, is applied to a continuous speech database. From the prediction error analysis, it is shown that word-level rules causes systematic errors at some supra-word level positions; pre-pause, phrase-final and sentence final. Furthermore, different durational characteristics between read and conversational sentences are shown.

In section 4, the word- and sentence- level factors are integrated into a single duration rule. Experimental results indicate the superiority of the rule combining both factors.

2 Analysis and modeling of word-level control

In this section, we discuss a formulation of word-level duration variation of Japanese segments using factor analysis. Since most variation in segment duration in Japanese occurs in the vowel [6], we focus on the vowel duration model in this section. The same analysis can be applied to consonants. For speech materials, 5,240 commonly used words uttered by a professional male narrator were used. The durations were measured by hand from spectrograms and entered into a database [14].

2.1 Factor analysis

As described in section 1, a number of factors and their effects on segmental duration have been described in previous works. However, there are few works in which these effects are compared quantitatively to establish a statistically satisfying duration control model. In this part we use a factor analysis method to determine quantitatively to what extent each factor affects durations.

The factors used in the analysis are those usually used for duration control, as shown in Table 1. They are taken into account in the analysis as follows. Corresponding to duration variance among phoneme types, the factor analysis was done for each vowels. Corresponding to the variance in location in the breath group, the vowels were separated into three groups; word initial, medial and final. C-V or V-C compensations can be described by the effects of neighboring phonemes. To take account of these effects properly, broad categories of neighboring phonemes were used. The phonemes are categorized mainly according to their manner of articulation, which is strongly associated with the inherent length of the segment, as shown in Table 2. As the number of moras in a breath group is the other dominant factor in duration control, the mora count of the word was also used for analysis.

Since a following geminate phoneme's effect on vowel length has not previously been studied, this effect was also examined.

To formulate the effect of these factors on vowel durations, Quantification Theory (type one) was used [15]. It is a kind of factor analysis, and it can formulate the relationship between categorical and numerical values in the form

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc} \delta_{fc}(i) \tag{1}$$

where, $\hat{y}_i$ is the predicted duration of the i-th sample, $\bar{y}$ is the mean duration of all samples and $\delta_{fc}(i)$ is the characteristic function

$$\delta_{fc}(i) = \begin{cases} 1 & \text{if i-th sample falls into category c of factor f} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

x_{fc} obtained by minimizing

$$(\hat{y}_i - y_i)^2 \tag{3}$$

correspond to regression coefficients of the usual linear regression model. We call this value "score" of the category.

In this way, the estimated length is computed as a linear combination of the following factors:

estimated length =
- mean length
- +preceding phoneme
- +following phoneme
- +mora count
- +following geminate consonant

As shown in this equation,

1. The wider the range of the score within the factor, the more dominant the factor is.
2. A negative(positive) score corresponds to the shortening(lengthening) effect of the category on duration.

2.2 Analysis results and word-level model

Figure 1 shows the analysis result for the vowel /i/ (1308 samples) in word-medial position. In this example, the minimized estimation error is 19.48 ms. Since the standard deviation of the duration distribution of word-medial /i/ is 32.49 ms, this model estimates 64% of the total variance of duration. In this figure, the score of each category, which represents the lengthening/shortening effects of category in millisecond, is illustrated.

The results show that the most important factor is the "preceding phoneme," the score of which ranges the most widely (99 ms). Each category's score for the factor "preceding phoneme" is inversely proportional to the average duration of the elemental phonemes. The score range for "following phoneme" is the second greatest and each category's score distribution is almost the same as for "preceding phoneme." These results show that the C-V and V-C compensations due to mora-timed characteristics are the dominant factors in duration control of Japanese. The range of the scores for the factor "word mora count" is 24.9 millisecond and is inversely proportional to the mora count of the breath group. This result shows that the mora count effect due to the distribution of breath is the second most important factor. From the analysis it can be seen that the effect of a following geminate consonant is not very large.

In Figure 2, the ranges of scores obtained by each analysis for each context are illustrated. As shown in this figure, it is confirmed that all analyses result in the same tendency as above, and that the durational characteristics of vowels, mentioned above, are common to all types of vowel and to all positions. Thus, this factor analysis clarifies the word-level control mechanisms quantitatively and yields a linear additive model of them.

Using these statistical values, segmental duration can be predicted by the same linear equation for all contextual conditions. For example, if an /i/ is preceded by /s/ and followed by /m/, in a four mora word and is not followed by geminate phoneme, the duration can be predicted as follows;

$$96.68 - 18.36 - 13.56 - 6.79 + 8.80 = 66.77\,ms$$

The terms of the left hand side of the equation correspond to; the mean duration, the effect of the preceding phoneme, the effect of the following phoneme, the effect of mora count and the effect of a following geminate phoneme.

2.3 Evaluation of the word-level model

The accuracy of the word-level control model was evaluated through an experiment on a 216 word database uttered by the same speaker which include 557 vowels. In the experiment, R.M.S. prediction error was 21.45 ms. This prediction error is smaller than the 32.57 ms standard deviation of all vowels in this set. Furthermore, for the restricted context used in our previous work [6], the estimation error is reduced to 18.3 ms which is smaller than 20.5 ms, the error reported in the same paper. Through these experiments, it is confirmed that this model can produce statistically reasonable duration using word-level controlling factors.

3 Analysis of sentence-level variance

The existence of sentence-level duration control mechanisms has been pointed out in earlier. However, sufficient statistical investigations based on a large database have never been reported. In this section, we apply the word-level duration model to sentences. Through error analysis of the word-level control model, the sentence-level control mechanisms are clarified statistically. Furthermore, using the result of this analysis, we propose a sentence-level control model.

For continuous speech database, conversational and read sentences of the same speaker as the isolated speech database were used. The total number of sentences is 282 as shown in the Table. 3. The duration of each phoneme was measured by the same method as for the isolated word database.

The sentence database consists of 7 categories; three conversational and four read sentences. The conversational categories were pronounced in three different manners. C1 and C2 categories consist of the same sentence but pronounced with pause at specified positions; after each long phrase(C1) and short phrase(C2), respectively. By specifying the location of pause in C1 and C2, we aimed at restricting the number of degrees of freedom in the production of conversational speech. The C3 category consists of the same sentences without any constraint.

Furthermore, to normalize differences in speaking rate, before error analysis, output durations of word-level model were shortened by the ratio of mean mora duration of isolated and continuous speech.

3.1 Duration setting experiment using word-level model

In Figure 2, average prediction errors for each category of samples are illustrated. In the figure, the R.M.S. error value shows the accuracy of the word-level model in predicting the vowel durations in the sentence. The mean error shows whether the word-level model over- or under-estimated the durations in the sentence. From the figure, it is clear that the segmental durations can not be predicted only from word-level effects. R.M.S errors of all categories exceed the average error in isolated speech, especially in conversational sentences.

A certain amount of negative mean error (under-estimation of durations) in each category results from the simple linear normalization of speaking rate.

3.2 Error analysis

3.2.1 Pre-pausal and Phrase-final lengthening

It is well known that segmental durations at word final position are longer than in word initial or medial position, in isolated word utterances of Japanese [6]. Figure 3 illustrates the duration variation of the vowel /a/ as a function of position in isolated word. This word final lengthening has been considered to be due to the same mechanism as pre-pausal lengthening in English. For this reason, the word-level rules are formulated separately, for word initial, medial and final positions.

In Figure 4, the average prediction errors of the segmental durations at pre-pausal, including phrase final, positions are illustrated. From the figure, it is clear that the word-level control model under-estimates the durations of pre-pausal vowels. In the figure, R.M.S prediction error values for pre-pausal are much greater than overall error because of the under estimation. The mean error of the model is about 20 milliseconds in read and 50 milliseconds in conversational sentences, respectively. This result implies that *pre-pausal lengthening* of Japanese should be treated as a different phenomenon from *word final lengthening*, especially in conversational speech.

Figure 5 illustrates the estimation errors in phrase final but not pre-pausal positions. In phrase final position, the errors are not as great as in pre-pausal position. This result shows that phrase final effect works as slightly shortening in both conversational and read sentences.

3.2.2 Sentence final effects

Figure 6 illustrates the errors in sentence final position. In sentence final position, segmental durations are over-estimated in read sentences and not so much under estimated as in pre-pausal position in conversational sentences. This phenomenon clearly shows that there are different duration control mechanisms in sentence final and pre-pausal positions. Furthermore, in read sentences, the word-level control model greatly over-estimates the durations. These results suggest that:

1. Different control mechanisms are at work for the segmental durations at *sentence final* and *pre-pausal* positions.

2. Different control mechanisms are at work for the segmental durations in *conversational* and *read* sentences at sentence final positions.

4 Sentence-level control modeling

In the previous section, three different durational phenomena associated with the location in the sentence are described. To show the durational manifestation of each phenomenon, the estimation errors at each position in read (D1) and conversational (C3) sentences are plotted in Figure 7. The figure shows that in conversational speech, the pre-pause effect is very important corresponding to a lengthening of about 40 milliseconds. On the other hand, in read speech, the sentence final effect is the most important, corresponding to a shortening of about 30 milliseconds. Based on these results, the combination of these duration variations and word-level model is expected to generate a more accurate duration control model than the conventional word-level model alone.

An experimental model combining word- and sentence-level rules was applied to the sentence data. The output of of the word-level model was adjusted as follows:

1. Lengthen pre-pausal segment durations by a factor of 1.55.

2. Shorten read sentence final segment durations by a factor of 0.63.

In this model, the two coefficients were determined by averaging the prediction errors from the above experiments. Figure 8 illustrates the overall estimation errors of this model. Estimation error of each category is decreased to the level of isolated word utterances except for freely uttered conversational sentences (C3). This

result reveals that these two controls are dominant in sentence-level duration control. The result also suggests the existence of some other control mechanisms for conversational speech.

5 Conclusions

In this paper we discussed the analysis and modeling of segmental duration in Japanese.

First, based on statistical analysis of an isolated speech database, we proposed a linear additive word-level duration rule. Through this modeling, following word-level effects were found:

1. The dominant factor for the word-level control is the effect of neighboring phonemes which comes from mora-timed characteristics of Japanese.

2. The second most important factor is the mora count of the breath group, due to the distribution of breath.

Second, the proposed rule was applied to sentential speech. The error analysis indicates that:

1. Pre-pausal lengthening in Japanese should be treated as a different phenomenon from word final lengthening.

2. A slight shortening effect occurs at phrase final positions.

3. Sentence final shortening occurs in read but not in conversational sentences.

Finally, an experiment using a combination of the word-level and sentence-level factors showed that the sentential effects were dominant in continuous speech. On this basis, we have proposed a new duration setting rule for speech synthesis. This rule is now used in the prosody control module of our synthesis system with satisfactory results.

We plan the following further research:

1. Analysis of the perceptual importance of the sentence-level factors.

2. Study of the effects of the rate of speech.

3. Analysis of more conversational speech.

6 Acknowledgments

The authors are grateful to Dr. Kurematsu for his encouragement and continuous support of our research and to Dr. Shikano for the discussion and suggestive comments.

References

[1] D.H.Klatt, "Review of text-to-speech conversion for English," J.Acoust.Soc.Am. 82, 760-761 (1987).

[2] N.Umeda, "Vowel duration in American English," J.Acoust.Soc.Am. 58, 434-445 (1975).

[3] S. Hiki, Y. Kanamori and J. Oizumi, "On the Duration of Phoneme in Running Speech"(in Japanese), Journal of Electronics, Information and Communication Engineers 50, 849-855 (1967).

[4] S. Hiki, "On the Duration of Various Segments in Sentence Speech"(in Japanese), Journal of Electronics, Information and Communication Engineers 50, 1465-1470 (1967).

[5] H. Sato, "Segmental duration and timing location in speech"(in Japanese), Trans. of the Committee on Speech Research Acoust.Soc.Jpn S77-31, (1977)

[6] Y. Sagisaka and Y.Tohkura, "Phoneme duration control for speech synthesis by rule"(in Japanese), Journal of Electronics, Information and Communication Engineers 67-A, 629-636 (1984)

[7] N.Higuchi and H.Fujisaki, "Durational control of segmental features in connected speech"(in Japanese),Trans. of the Committee on Speech Research Acoust.Soc.Jpn S80-40, (1980)

[8] H.Kawasaki, "Models and data on the temporal regulation of speech: Isochrony in Japanese and English"(in Japanese), J.Acoust.Soc.Jpn., 39, 389-397 (1983)

[9] R. F. Port, J. Dalby and M. O'Dell, "Evidence for mora timing in Japanese," J.Acoust.Soc.Am., 81, 1574-1585 (1987)

[10] D.H.Klatt, "Linguistic uses of segmental duration in English: acoustic and perceptual evidence," J.Acoust.Soc.Am., 59, 1208-1221 (1976)

[11] Lehiste I., "Some temporal aspects of spoken discourse," Speech Communication Papers Presented at the 97th Meeting of ASA.

[12] B. Rackerd, W.Sennett and C.A.Fowler, "Domain-final lengthening and foot-level shortening in spoken English," Phonetica 44, 147-155, 1987

[13] W.N.Campbell "Extracting speech-rate values from a read-speech database," Proc. ICASSP 88, (1988)

[14] K.Takeda, Y.Sagisaka, S.Katagiri and H.Kuwabara "A Japanese speech database for various kinds of research purposes"(in Japanese), J.Acoust.Soc.Jpn., 44, 747-754 (1988)

[15] C. Hayashi, "Recent Theoretical and methodological developments in multidimensional scaling and its related methods in Japan," Behaviormetrika, 18

factor	acoustic manifestation	motivation
type of phonemes	inherent length compressibility	articulatory constraint
neighboring phoneme	C-V compensation	mora-timing
mora number location in a breath group	word final lengthening	preset timing associate with breathing
speaking rate	average duration	speaking tempo

Table 1: Control factors for Japanese segmental durations

factor	categories
neighboring phoneme	(s,sh,ts,ch),(p,t,k),(h,f),(m,n), (j,z),(w,y),(b,d,g),(r), (a),(i),(u),(e),(o),(N), (palatalized vowel like portion)
mora in a breath group	(1),(2),(3),(4),(5),(more than 5)
followed by geminate	(yes),(no)

Table 2: Categories for each factor

set	task	sentences	uttering condition	vowels
C1	conversational	50	pause after every long phrase	1248
C2	conversational	50	pause after every short phrase	1199
C3	conversational	50	not specified	1230
D1	read	22	not specified	816
D2	read	22	not specified	779
D3	read	43	not specified	1373
D4	read	45	not specified	1372

Table 3: Sentence database

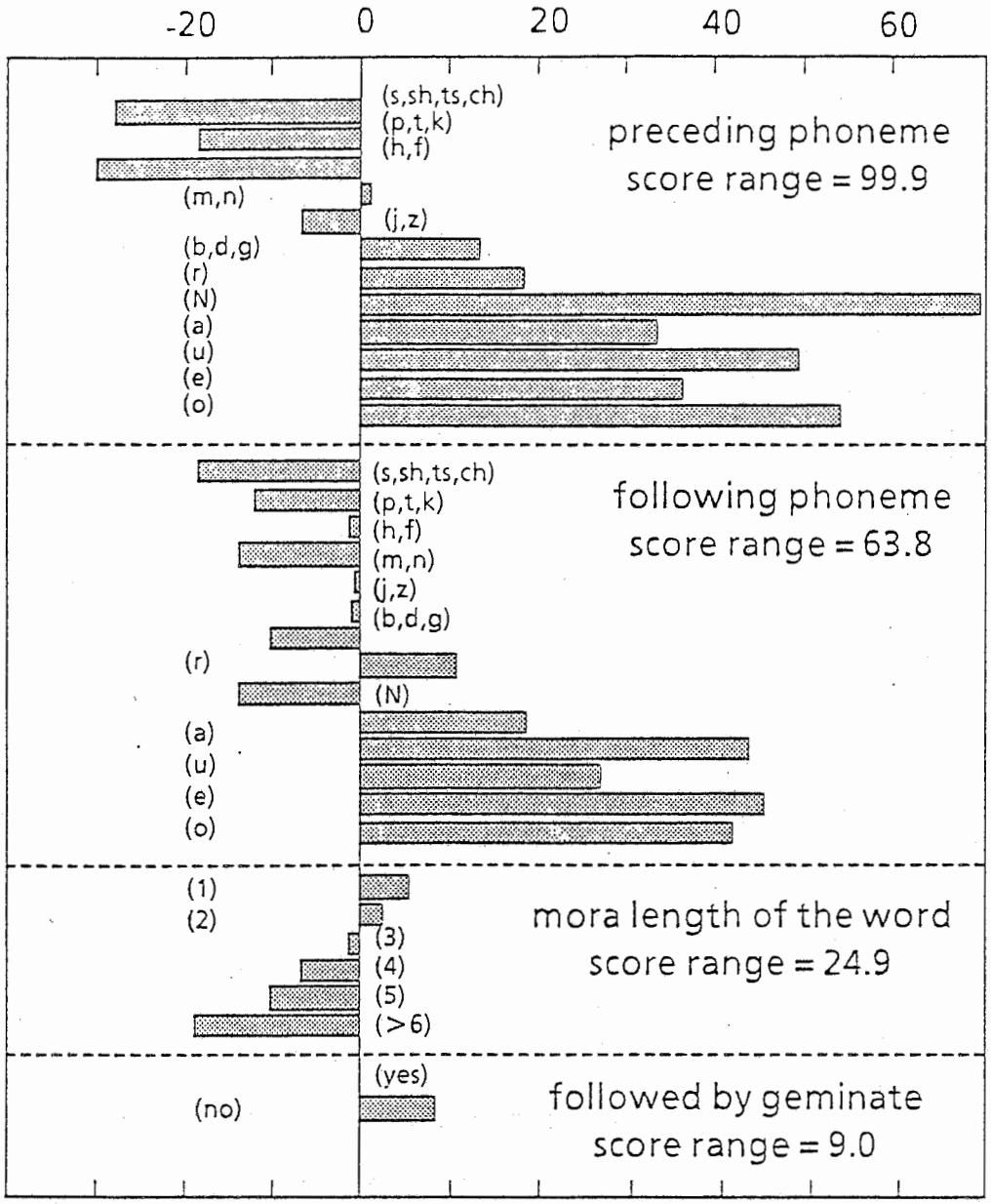


Figure 1 An example of analysis (word medial /i/)

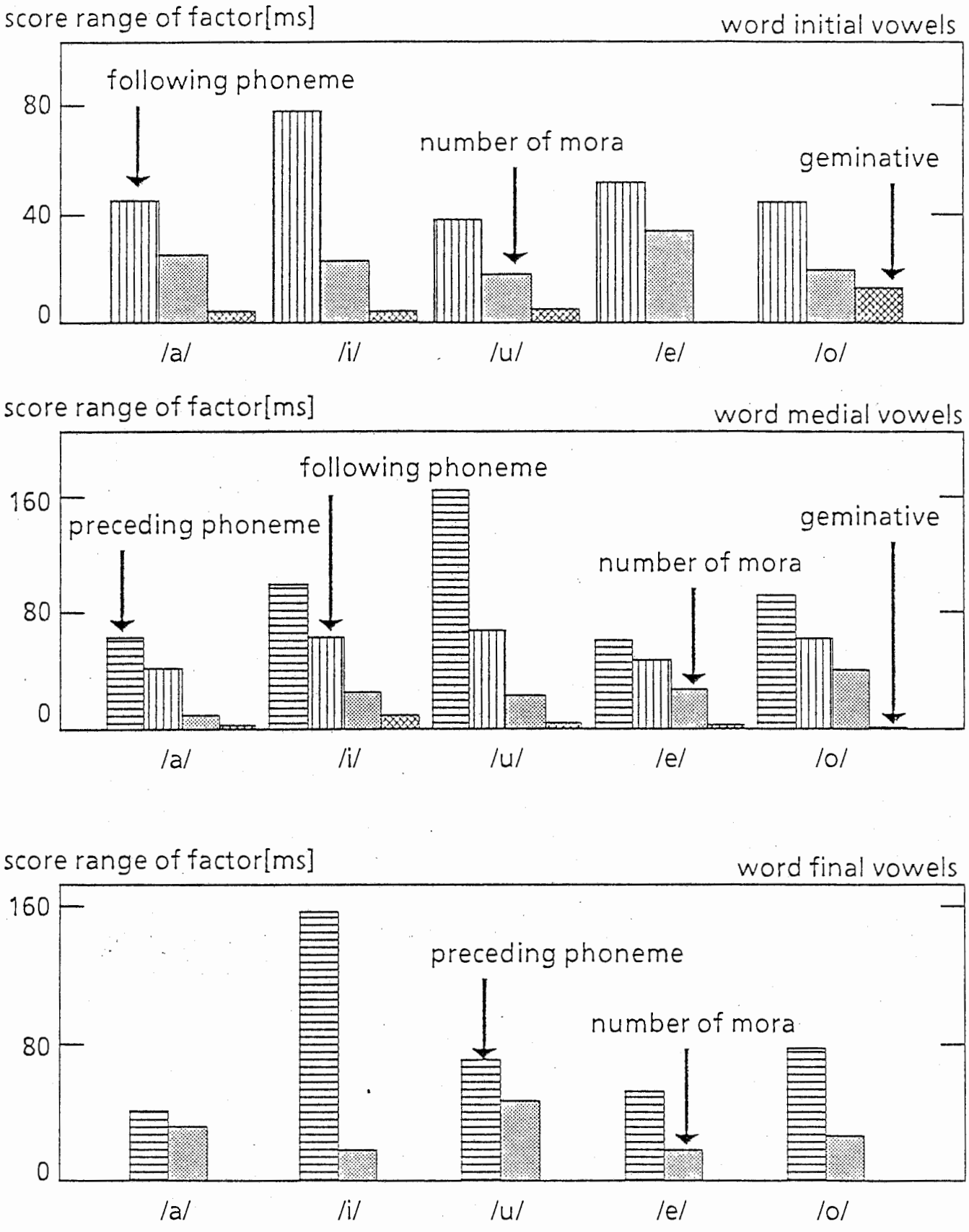


Figure 2. Results of factor analysis

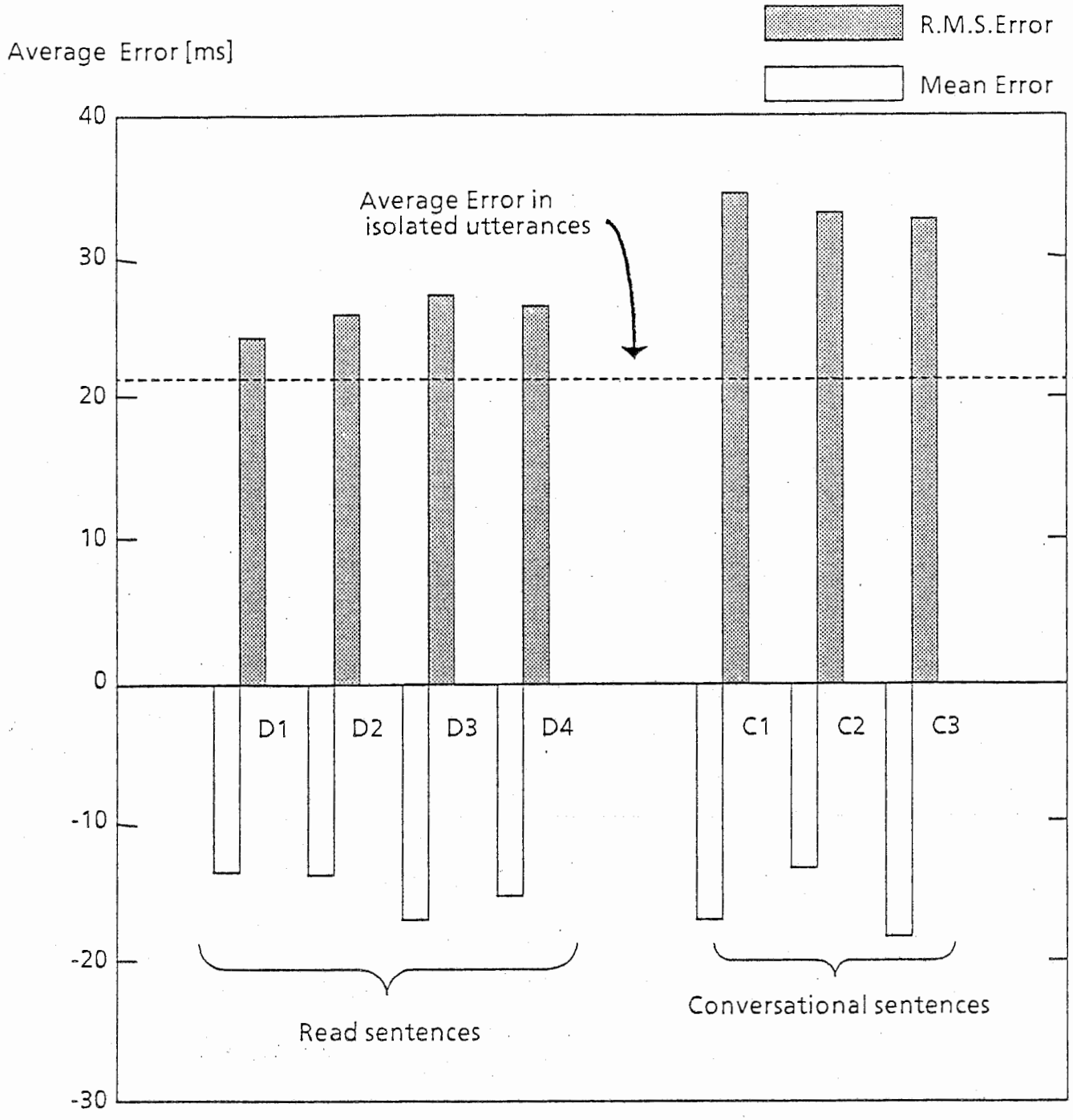


Figure 3. Overall error in continuous speech

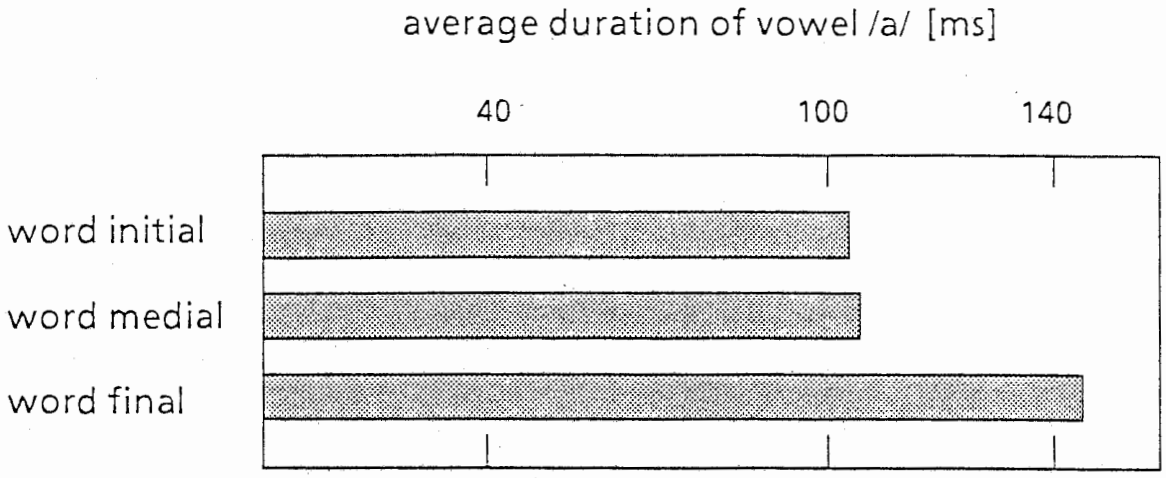


Figure 4. Word final lengthening

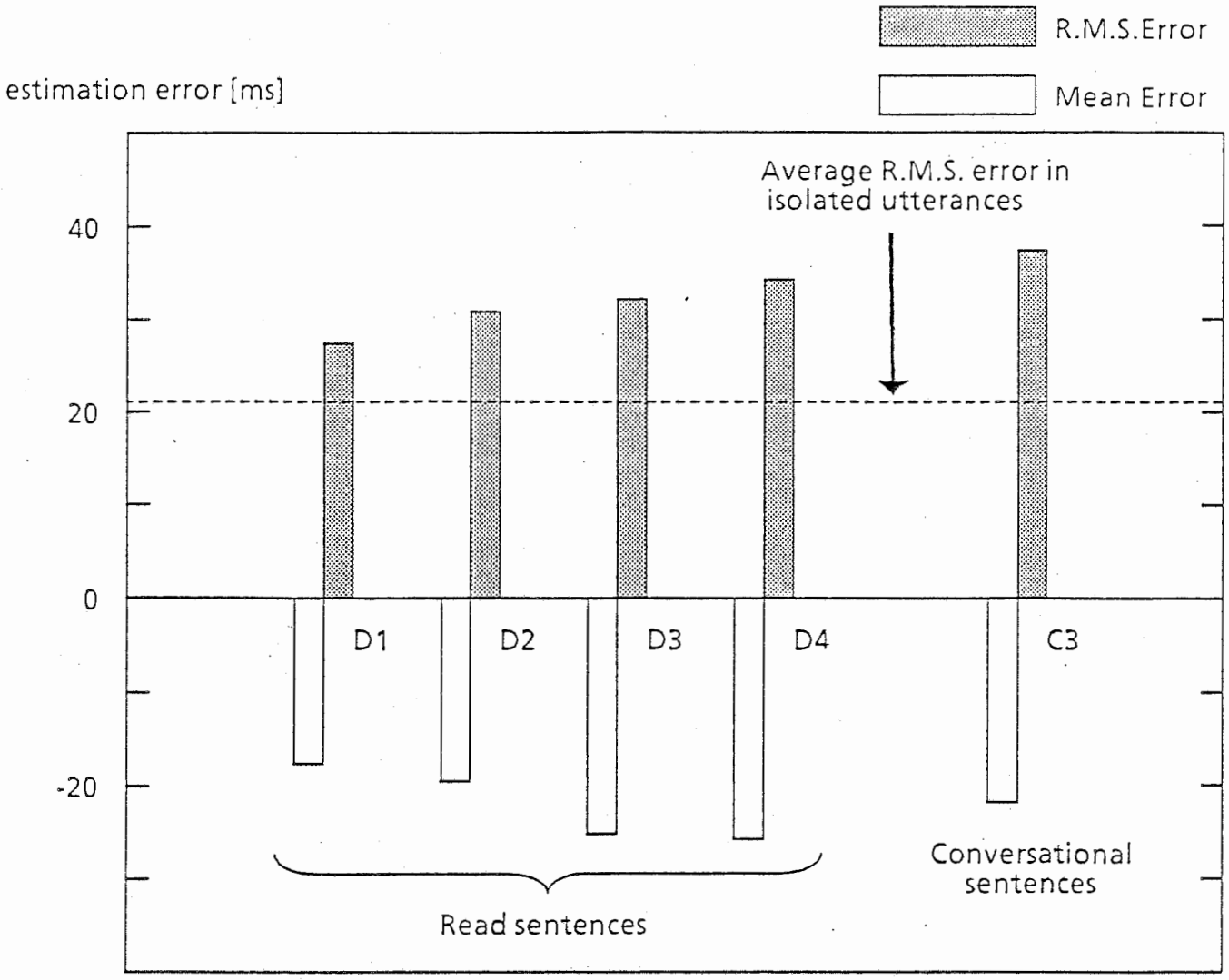


Figure 6. Error at phrase final position

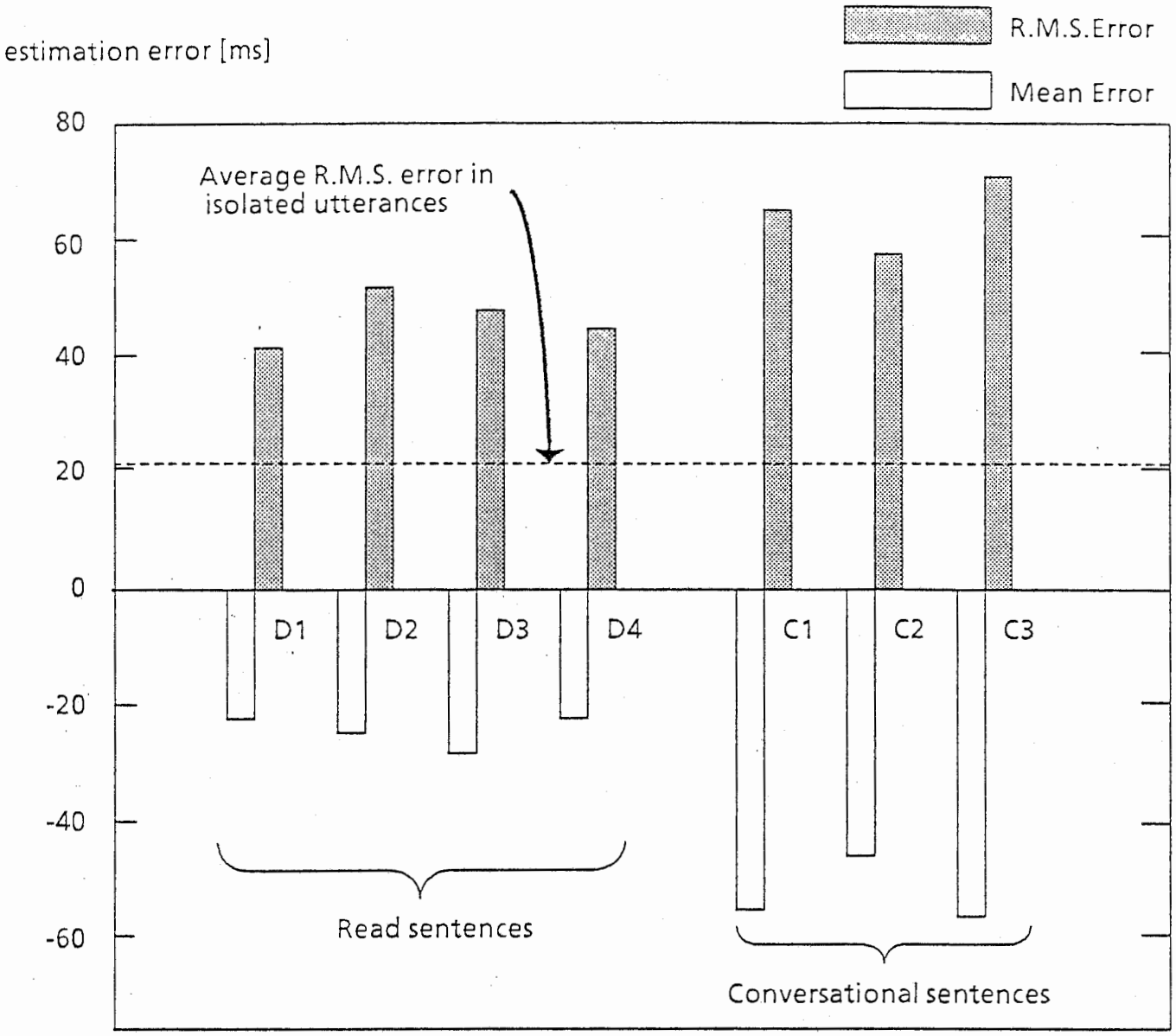


Figure 5. Errors in pre-pausal position

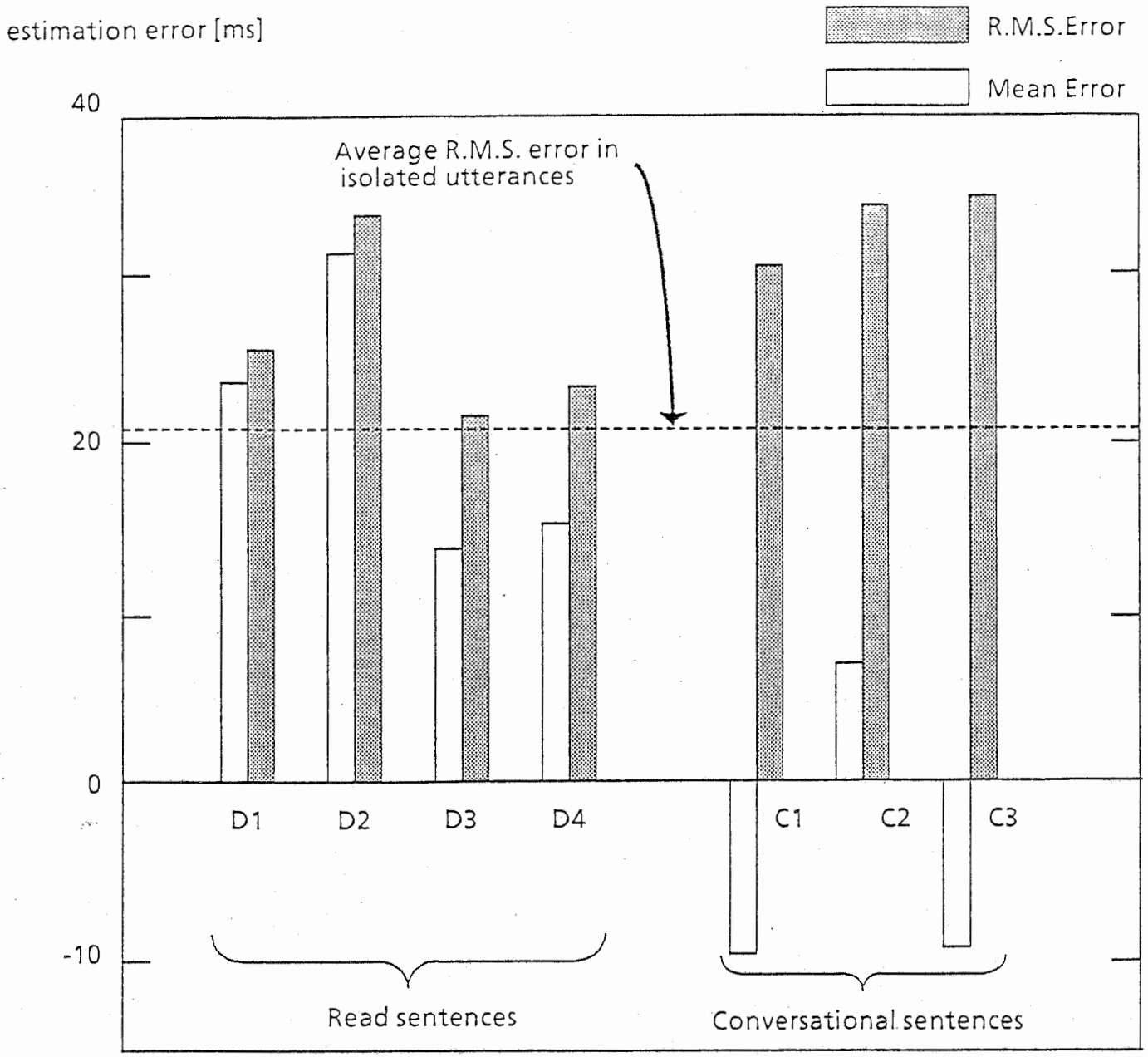
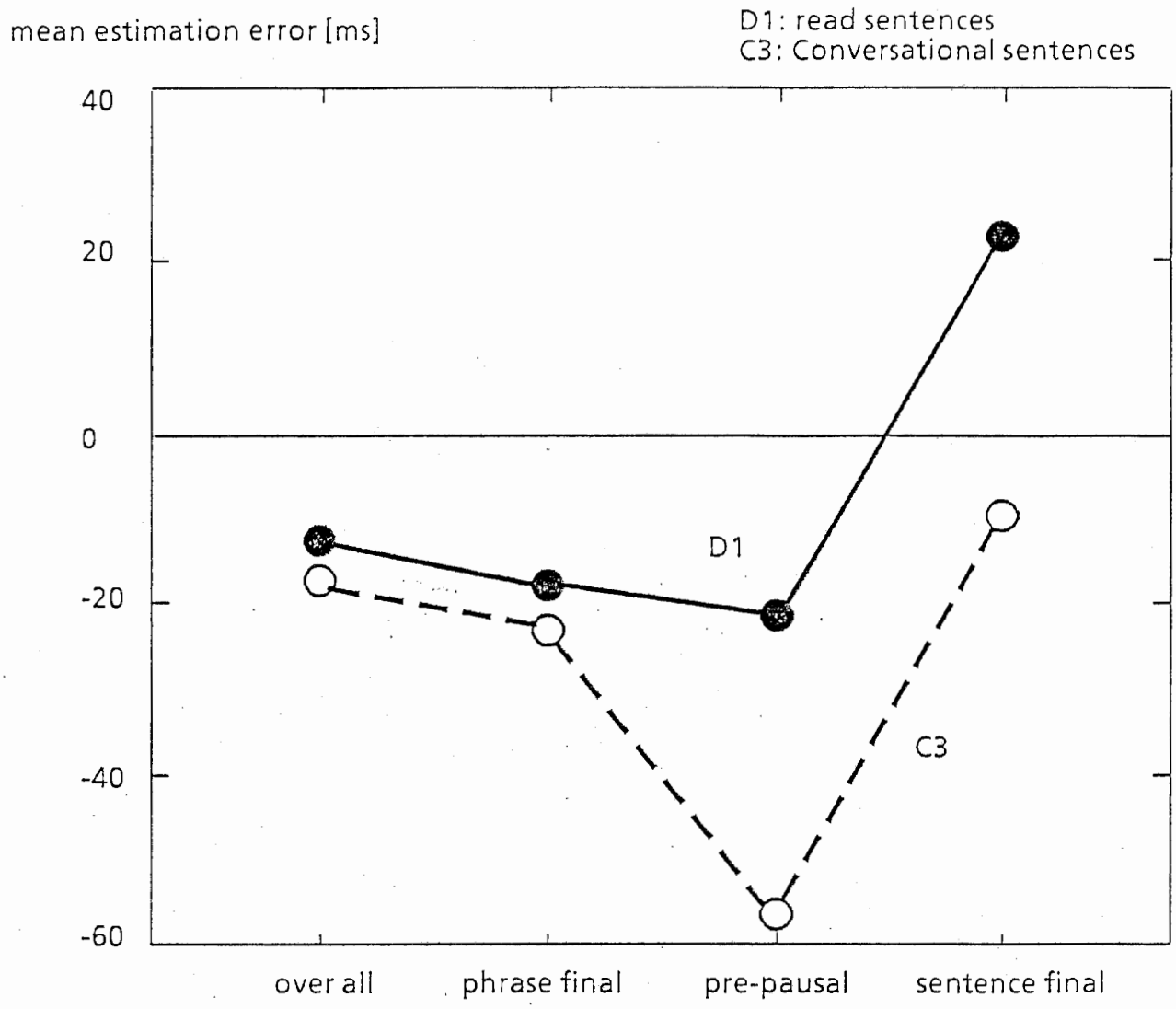


Figure 7. Errors in sentence final position



Figur 8: Estimation error in various positions

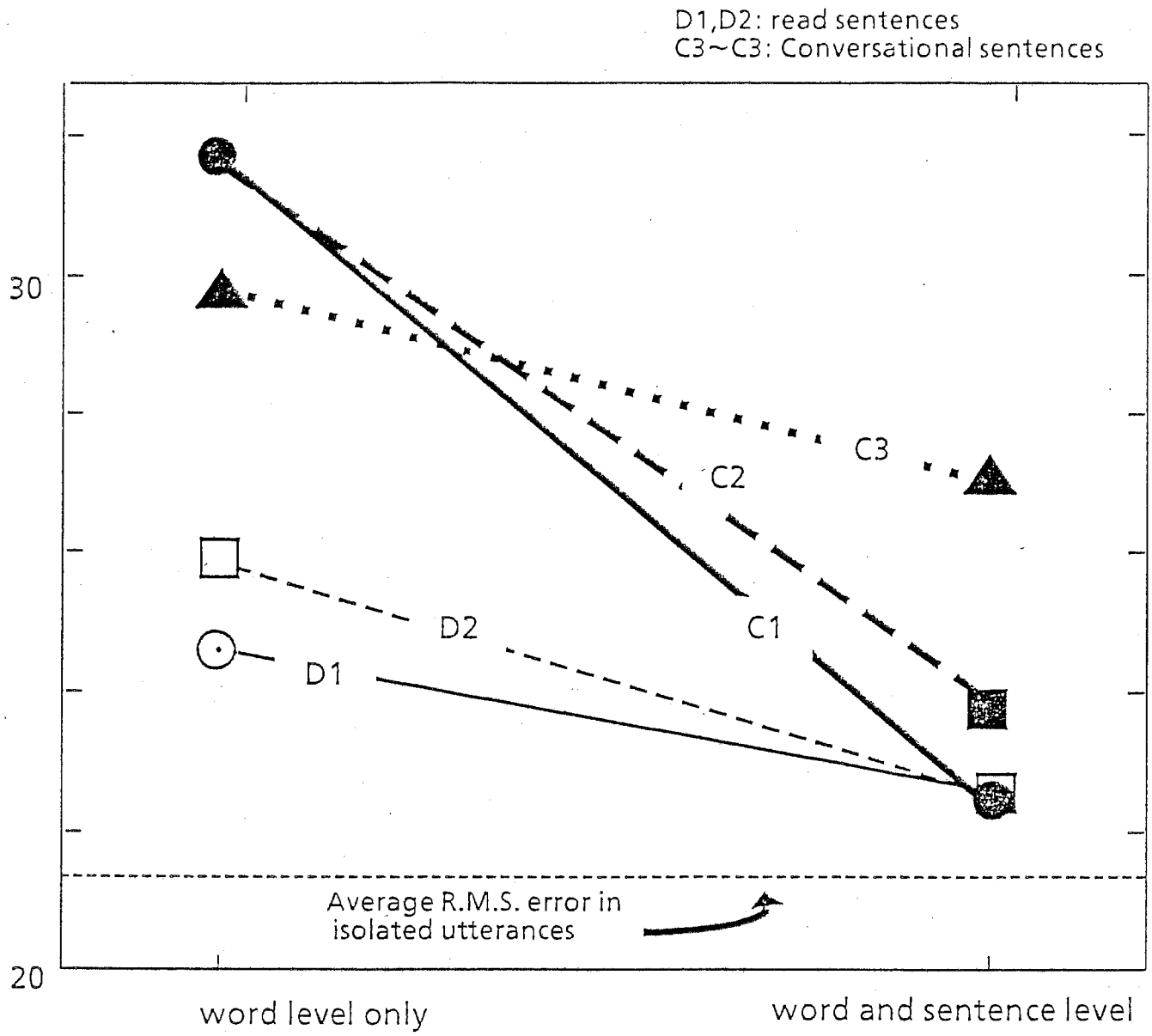


Figure 9. Combing word and sentence level effects