Relationship between control precision and perceptual sensitivity to segmental duration

Hiroaki Kato*, Makiko Muto‡*, Minoru Tsuzaki†, and Yoshinori Sagisaka‡†

*ATR Human Information Science Laboratories, Kyoto 619-0288, Japan

†ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan

‡GITI, Waseda University, Tokyo 169-0051, Japan

E-mail: kato@atr.co.jp, makiko.muto@ruri.waseda.jp, minoru.tsuzaki@atr.co.jp, sagisaka@giti.waseda.ac.jp

ABSTRACT

This paper investigates a potential linkage between production and perception in the temporal aspect of speech from an ecological point of view. The variability of the segment duration is analyzed to estimate the temporal control precision in speech production, using large-scale speech corpora comprising a commonly used word set (5240 word entries spoken by 20 speakers) and a phonetically balanced sentence set (503 sentence entries spoken by 10 speakers). The results first show that the duration of a given segment has little correlation with the estimated control precision of that segment. They also show that the estimated precision systematically changes as a function of vowel quality and temporal position in an utterance unit, such as a word or phrase. These findings are in good agreement with the perceptual characteristics, i.e., sensitivity and acceptability of deviation in segment duration; accordingly, the results suggest the existence of some linkage between perception and production.

1. INTRODUCTION

To estimate the temporal precision required for the duration rules of speech synthesis applications, studies have investigated perceptual sensitivity to temporal changes in speech segments [6, 1]. Along these lines, a series of experiments has systematically revealed that human perceptual sensitivity or acceptability of temporal deviation in segment duration is significantly affected by several segmental and contextual properties of the segment in question, e.g., the phoneme class (vowel, nasal, or fricative) and temporal position within an utterance unit such as a word or phrase [3, 4, 5, 10].

The observed perceptual effects have been, in general, accounted for by using the psychoacoustical characteristics of the speech material itself. For instance, a strong correlation has been found between these perceptual effects and psychoacoustic properties such as the loudness of the segment in question. However, from an ecological point of view, one's perceptual characteristics develop under the strong influence of the exposed environment during the process of growth. Thus, one's perceptual acceptability of a temporal change in a segment at a particular context may reflect the temporal deviation generally expected for a segment at that context in his/her environment.

On the other hand, the control precision or strategy of a speaker may reflect the perceptual characteristics as a result of the feedback through his/her auditory system. However, it is still an open question whether such context-dependency of perceptual sensitivity is also consistent with the temporal properties of spoken utterances, i.e., whether the control precision has correlation with the perceptual sensitivity.

Investigating this possibility could yield useful information for understanding the nature of each of the perceptual factors found, whether it involves an ecological constraint during the development process or solely reflects inherent auditory characteristics. Such analysis could provide useful information not only for exploring the ecological influence on the perceptual factors but also for estimating the requisite precision of durational rules in speech synthesis.

To answer this question, the present study, as a first step, precisely analyzed the context-dependency of the control precision, or variability of segmental durations, by using large-scale speech corpora. First, in Section 2, the major effects of the segment attribute and context on the perceptual sensitivity (or acceptability) to deviations in the segment duration are summarized from previous studies. Second, in Section 3, the methodology of estimating the control precision is described. Finally, Section 4 shows the results of the data analysis and discusses the relationship between control precision and perceptual sensitivity.

2. PERCEPTUAL EVALUATION OF DEVIATION IN SEGMENT DURATION

2.1. Validity of segmental duration as a basis of perceptual evaluation

This first subsection examines the relationship between a given segmental duration and the range of acceptable or detectable temporal deviation of that segment. Kato *et al.* found that the acceptable range in milliseconds does not correlate with the segmental duration ($r = 0.02$) [4]. A similar lack of relevance was observed in the sensitivity study as a non-correlation of the just noticeable difference with its base duration [7]. These results suggest that the listeners do not evaluate the segmental deviation in relation only to the base segmental duration itself.

2.2. Segmental and contextual factors affecting perceptual evaluation

Both the attributes and contexts of a segment are known to affect the perceptual evaluation of temporal deviations of that segment. The former includes the phonemic type of the segment in question. The deviation of a vowel duration is detected more easily than that of a consonant duration [1, 5]. Among vowels, a change in /a/ is perceived as more salient than a change in /i/. Figure 1 (a) illustrates the effect of vowel quality on the acceptability of the deviations in

ISBN 1-876346-48-5

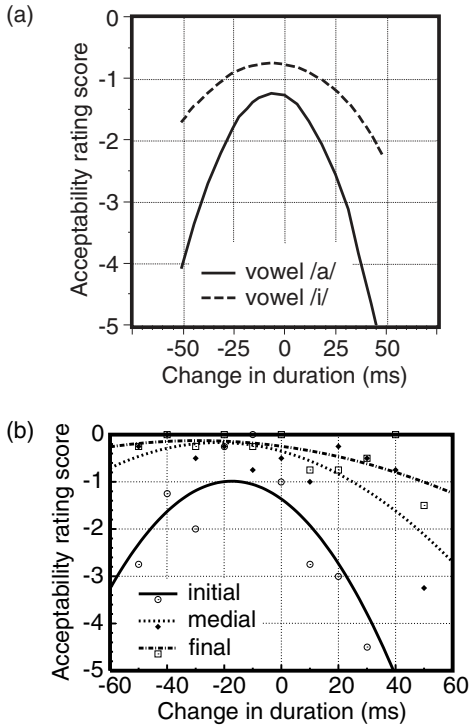

Figure 1: Differences in perceptual sensitivities to segment duration. (a) Effect of vowel quality. An example showing the difference in listeners' evaluation between vowels /a/ and /i/ in word speech. (b) Effect of temporal position in an utterance unit. An example showing the difference in listeners' evaluation between initial, medial, and final positions in the phrase /tonikaku/ embedded in a sentence. The parabolic fitting lines are shown in both panels.

segment duration. The acceptability rating score decreases more drastically for /a/ than for /i/ with the duration change. Such a phonemic effect can be accounted for by the loudness value of the deviated segment [5], i.e., a segment with a larger loudness is more perceptually salient than that with a smaller loudness.

The latter, a contextual factor, includes the effects of the temporal position in an utterance unit and the type of adjacent segments. It was found that the deviation of a word-initial segment is more salient than that of a word-medial segment [4]. A similar tendency has been found as the intra-phrase positional effect as shown in Fig. 1 (b): The listeners were most sensitive to a phrase-initial deviation and least sensitive to a phrase-final one, with phrase-medial deviation having an intermediate sensitivity [10]. With regard to the type of adjacent segments, a duration change of a segment was more saliently perceived when it was followed by an unvoiced segment than when it was followed by a voiced segment [4].

3. MEASURE OF CONTROL PRECISION

3.1. Speech corpora

The source speech data were taken from the ATR Japanese speech database [8]; they were a set of commonly used words spoken by 10 male and 10 female professional speakers (5240 words/speaker, 104800 words in total) and a phonetically balanced set of sentences spoken by 6 male and 4 female professional speakers (503 sentences/speaker, 5030 sentences (113220 words) in total).

They were segmented by trained transcribers with a 2.5-ms precision. The average segmentation disagreement between transcribers ranged from 0.5 to 7.5 ms irrespectively of the difference in phoneme. This range is sufficiently smaller than the expected range of the control precision to be studied in this paper.

3.2. Cancellation of known control factors

The variation in segment duration of the prepared data includes several systematic deviations in addition to those corresponding to the control precision. They are the deviations intrinsic to particular segmental or contextual properties caused by the control factors of speech production [12, 2]. The influence of these factors can be observed in intrinsic duration, durational compensation among neiboring phonemes, lengthening or shortening associated with the difference in the temporal position within an utterance unit, and other effects; the factors are summarized in Table 1 of Sagisaka's paper in this volume [11].

To cancel out the influence of these control factors as much as possible, a linear regression model was applied to quantize the influence on the segmental durations by each of the control factors. Next, segmental durations were synthesized to replicate the observed ones by using the quantized factors in accordance with the linear prediction procedures [2]. Finally, the influence of the control factors was cancelled out by subtracting the predicted segmental durations from the corresponding observed ones. These residual errors were regarded as representing the control precision in this study because each error corresponded to the deviation of each observed duration from the expected duration for that particular phoneme and context.

The control factors considered were basically the same as those of the previous study [2]. These cancelling procedures were applied to the data from each speaker. The multiple correlation coefficients which represent the goodness of each linear regression, ranged from 0.734 to 0.866 and were comparable with or better than those in the previous study. Figure 2 shows the histograms of the segmental durations and the residual errors (the data from the sentence speech of all ten speakers).

Note that the normal distribution function accurately traced the histogram of the residual errors as superimposed in the lower panel of Fig. 2. This paper, therefore, took this characteristics into account and exploited the standard deviation as an index representing the width of error distribution or the control precision.

4. RELATIONSHIP BETWEEN CONTROL AND PERCEPTUAL CHARACTERISTICS

This section shows the estimated control precision under several conditions and discusses their relevance to perceptual sensitivity, with a focus on the vowel segments in the sentence speech material. This is because the control of vowel duration is in general more flexible than that of consonant duration [6] and, therefore, a large amount of temporal variation could be expected for it; the sentence material was phonetically balanced, so the segments adjacent to the vowels are expected to maintain a maximum phonetical vari-

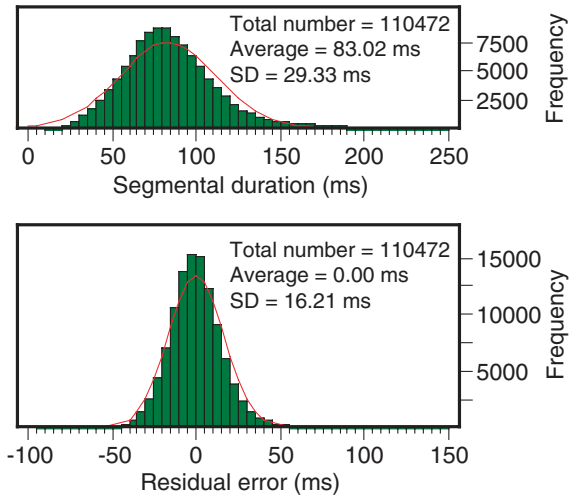

Figure 2: The raw and normalized distributions of segmental duration. Upper: The histogram of the observed segmental duration with a fitting curve of the normal distribution. Lower: The histogram of the residual error, i.e., a predicted duration minus the corresponding observed duration, with a fitting curve of the normal distribution.

ation.

4.1. Validity of segmental duration as a basis of control precision

Figure 3 shows the absolute values of the residual errors as a function of the corresponding original segmental duration. Although a notable positive correlation can be observed in the region where the segmental duration is 130 ms or longer, the percentage of samples in this region is only 8%. On the other hand, no correlation was observed ($r = 0.0058$) in the remaining region, which includes 92% of the samples. This lack of correlation between the original segment duration and the control precision is in good agreement with the lack of correlation between the original segment duration and the range of acceptable deviations observed in the perceptual studies.

The remaining part of this paper, therefore, does not employ relative errors to original durations but the raw errors (in ms) in referring to the control precision.

4.2. Vowel quality, position in an utterance unit, and voicing of adjacent segments

This subsection examines the phonemic and contextual dependencies of the control precision using the standard deviation of the residual errors.

First, the effect of vowel quality was examined. Figure 4 shows control precision as a function of the vowel quality of the segment in question. Each point represents the observation from each speaker. The estimated control precision was worst for vowel /i/ and best for vowels /a/, /e/, /o/, with vowel /u/ between those. A similar tendency was also observed in the word speech, i.e., the control precision of high vowels tended to be worse than that of mid or low vowels. This tendency is in good agreement with the perceptual characteristics introduced in Fig. 1 (a). Note that the effect of vowel quality on the perceptual sensitivity could be accounted for by the loudness of the segment in

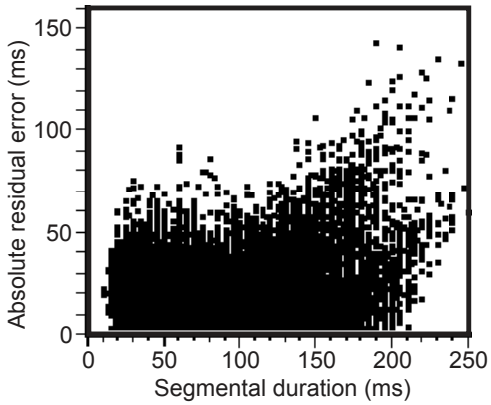

Figure 3: Absolute prediction error as a function of segment duration. Although a mild positive correlation is observed ($r = 0.24$) as a whole, there is no correlation ($r = 0.006$) if the x-axis is limited up to 130 ms; this range covers more than 90% of the observed points.

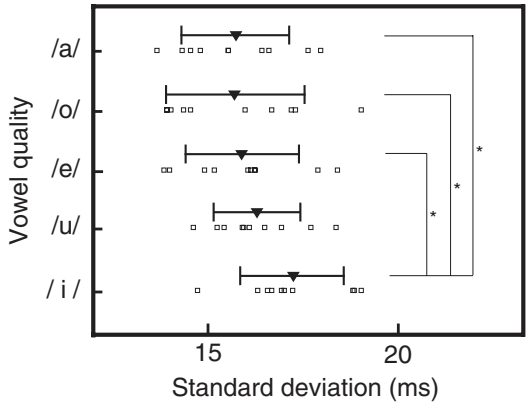

Figure 4: Vowel control precision represented by the standard deviation (SD) of the residual errors as a function of vowel quality. Each point represents the SD of each of ten speakers. Triangles and error bars show the averages and standard deviations of the SDs, respectively. Asterisks mark the pairs of levels whose averages are significantly different from each other ($p < 0.05$).

question; a larger loudness corresponded to a smaller acceptable deviation. This also seems to be the case also in the control precision; it is acknowledged that the loudness of high vowels is generally smaller than that of mid or low vowels [9].

Second, the effect of temporal position in an utterance unit was examined. As shown in Fig. 5, the control precision was worse at the final position of an utterance unit than at the initial or medial position. This tendency was commonly observed for all units examined, i.e., a word, minor phrase, accent phrase, breath group, and sentence. This tendency is in good agreement with the perceptual characteristics introduced in Fig. 1 (b), with the exception of that between initial and medial positions, where the perceptual effect was observed while the effect on control precision was not always observed.

Finally, the effect of the adjacent segment was examined. Although the perceptual sensitivity was higher for a segment followed by an unvoiced segment than for one fol-

ISBN 1-876346-48-5

lowed by a voiced segment, no significant difference was observed in the control precision due to the voicing of the adjacent segment.

5. CONCLUSIONS

For clues to a linkage between control precision and perceptual sensitivity, a comparison was performed between the variability of the segmental durations in spoken utterances and the perceptual evaluation of deviations in segment duration. The results showed that several segmental or contextual factors, such as the vowel quality and temporal position in an utterance unit, commonly affected both production and perception in a similar way. On the other hand, some of the other factors affecting the perceptual evaluation had apparently no effect on control precision; these included the voicing of the following segment and the difference in the initial and intermediate intra-unit positions. Clearly, a linkage is not necessarily found in every aspect between perception and production. However, further studies are expected to determine the crucial conditions that are responsible for the existence of agreement between the control precision and perceptual sensitivity.

ACKNOWLEDGMENTS

This work was supported in part by a contract with the Telecommunications Advancement Organization of Japan.

REFERENCES

[1] R. Carlson and B. Granström. Perception of segmental duration. In A. Cohen and S. Nooteboom (eds.), *Structure and Process in Speech Perception*, pp. 90–106. Springer, Berlin, 1975.

[2] N. Kaiki and Y. Sagisaka. The control of segmental duration in speech synthesis using statistical methods. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure*, pp. 391–402. IOS, Amsterdam, 1992.

[3] H. Kato, M. Tsuzaki, and Y. Sagisaka. Acceptability for temporal modification of consecutive segments in isolated words. *J. Acoust. Soc. Am.*, 101:2311–2322, 1997.

[4] H. Kato, M. Tsuzaki, and Y. Sagisaka. Acceptability for temporal modification of single vowel segments in isolated words. *J. Acoust. Soc. Am.*, 104:540–549, 1998.

[5] H. Kato, M. Tsuzaki, and Y. Sagisaka. Effects of phoneme class and duration on the acceptability of modifications in speech. *J. Acoust. Soc. Am.*, 111:387–400, 2002.

[6] D. H. Klatt. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am.*, 59:1208–1221, 1976.

[7] D. H. Klatt and W. E. Cooper. Perception of segment duration in sentence contexts. In A. Cohen and S. G. Nooteboom (eds.), *Structure and Process in Speech Perception*, pp. 69–89. Springer, Berlin, 1975.

[8] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Commun.*, 9:357–363, 1990.

[9] I. Lehiste. *Suprasegmentals*. MIT, Cambridge, 1970.

[10] M. Muto, H. Kato, M. Tsuzaki, and Y. Sagisaka. Effects of intra-phrase position on acceptability of change in segmental duration in sentence speech. In *Proc. 7th ICSLP*, pp. 761–764, 2002.

[11] Y. Sagisaka. Modeling and perception of temporal characteristics in speech. In *Proc. 15th ICPhS*, 2003. in this volume.

[12] Y. Sagisaka and Y. Tohkura. Phoneme duration control for speech synthesis by rule. *Trans. Inst. Electron. Commun. Eng. Jpn.*, J67-A:629–636, 1984. in Japanese with English figure captions.

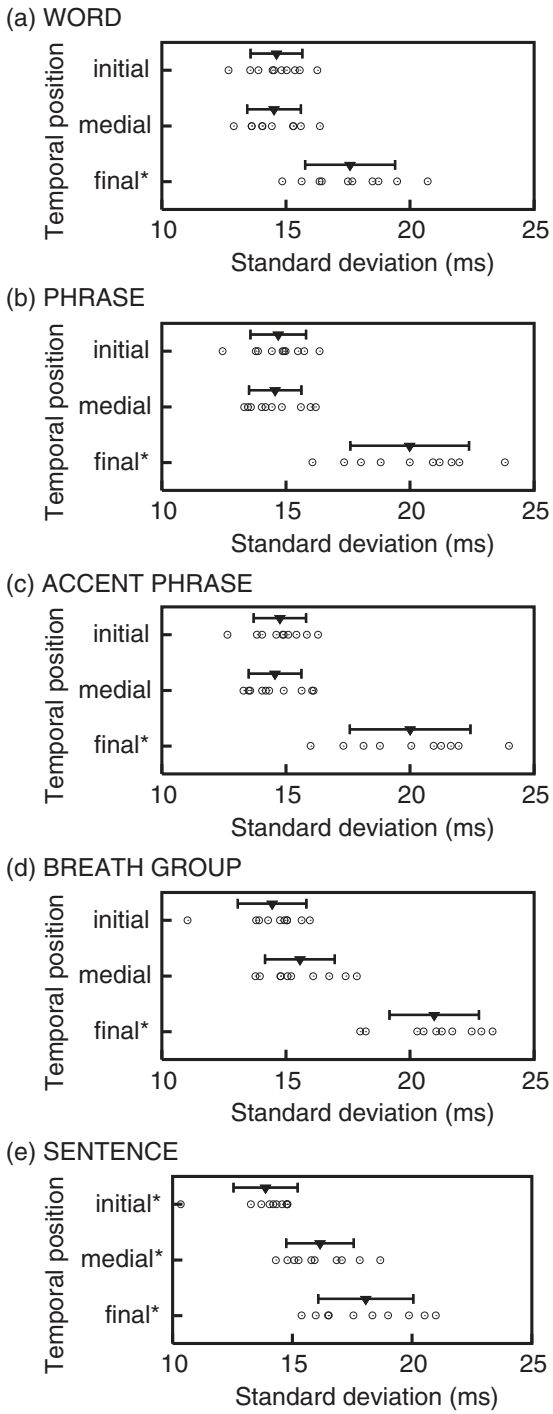

Figure 5: Vowel control precision represented by the standard deviation (SD) of the residual errors as a function of the temporal position in the following utterance units: (a) Word, (b) Phrase, (c) Accent phrase, (d) Breath group, and (e) Sentence. Each point represents the SD of each of ten speakers. Triangles and error bars show the averages and standard deviations of the SDs, respectively. Asterisks mark the levels whose averages are significantly different from those of the other levels ($p < 0.05$).