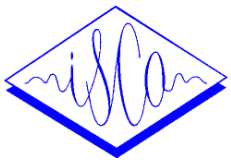

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ACCEPTABILITY OF TEMPORAL MODIFICATION IN CONSONANT AND VOWEL ONSETS

Hiroaki KATO*, Minoru TSUZAKI*, and Yoshinori SAGISAKA**

*ATR Human Information Processing Research Laboratories
**ATR Interpreting Telecommunications Research Laboratories
Hikaridai, Seikacho, Kyoto, 619-02 Japan
E-mail:kato@hip.atr.co.jp

ABSTRACT

To specify the location of perceptually dominant markers for temporal structures of speech, the acceptability or detectability of the modification of segmental duration is measured. The modification is carried out in a complementary way, i.e., two successive segments are lengthened or shortened, and have the same absolute duration and opposite directions of change. The first experiment using 15 four-mora word stimuli shows that a vowel (V) duration and its adjacent consonant (C) duration can perceptually compensate each other. This compensation is found to not depend on the temporal order of target pairs (C-to-V or V-to-C), but rather on the loudness difference between V and C; the acceptability decreases when the loudness difference between V and C becomes high. This suggests that perceptually dominant markers locate around major loudness jumps; this finding is supported by the second experiment using non-speech stimuli replicating the loudness contours of speech stimuli. The detectability of temporal displacement is higher at large loudness jumps than at small loudness jumps.

I. INTRODUCTION

As temporal structures such as rhythm or tempo can be perceived in speech, there should be some kinds of markers giving us such temporal information about speech. These markers, of course, will in all likelihood have temporal positions that change when the duration of a single segment is modified. As such, when two or more modifications are made, they may cancel each other out, and consequently some of the markers might become unable to move. If, however, such a temporal compensation can be made to preserve perceptually dominant markers, then the perceived temporal structure will suffer little from multiple modifications. Thus, in this case the perceptual compensation effect can be used as an index for candidates of the perceptually dominant temporal markers. The purpose of this study is to specify such dominant markers in speech sounds by measuring the perceptual compensation effect and to investigate which features the perceptually dominant markers are based on, i.e., linguistic features or acoustic ones.

Several studies have looked at the perception of temporal modifications for speech segments[1, 2], but only a few studies have addressed perceptual phenomena caused by interactions among multiple modifications. For instance, Hoshino and Fujisaki[3] investigated the human's tolerability of changes in durations of vowel (V), consonant (C), C-to-V, or V-to-C. They summarized that the tolerability is higher when durational changes of a C and its succeeding V complement each other. This suggests that perceptually dominant markers are likely to locate around C onsets. In the stage of speech production, the compensatory relation between a C duration and its succeeding V duration was also observed in studies that performed acoustical analyses on a large database of spoken Japanese[4, 5].

These previous studies apparently support the hypothesis that the compensation effect at the level of CV or VC is governed by a unit comprising a C and its succeeding V, in both the perception and production stages. This hypothesis is likely to be supported from linguistic considerations because a CV unit usually coincides with a mora, a phonological segmentation unit. We refer to this hypothesis as *the CV hypothesis* hereafter.

However, Sato has provided evidence against *the CV hypothesis*[6, 7]. He found that the loss of acceptability caused by the lengthening of the first or the third vowel in the word *sakanayasan* (a fish dealer) could be compensated more by the shortening of its succeeding consonant than by the shortening of its preceding consonant. He also reported a VC compensation in speech production based on acoustical analyses as well as a CV compensation. Thus, it is still an open question whether CV is more significant for perceptual compensation than VC; *the CV hypothesis* therefore needs to be tested.

Even a psychophysical "non-speech" study observed a perceptual compensation effect; Schulze[8] found that complementary temporal modifications on two successive intervals were more difficult to detect than a modification on a single interval in the detection of temporal displacement within seven successive intervals. This finding clearly indicates the necessity of inducing a non-linguistic (probably acoustic) point of view as well as a linguistic one to understand the nature of the temporal compensation effect.

In the current study, we therefore induce two hypotheses on the effect of temporal compensation; each of them being based on a linguistic or acoustic context. The first one is *the CV hypothesis*, which is tested in the first experiment measuring acceptability of durational modifications using thirty pairs of V and C within fifteen four-mora word stimuli (CVCVCVCV). This hypothesis will be supported if a complementary modification of a C and its succeeding V is generally more acceptable than that of a V and its succeeding C.

The second hypothesis focuses on the role of loudness-related contexts from among various acoustic cues. This is because previous studies[9-11] showed that perceptual sensitivity to durational modification on a single segment is highly affected by the loudness of the target segment and that of its adjacent segments. The second hypothesis, *the loudness hypothesis*, assumes the loudness difference at V-to-C or C-to-V transition as the dominant factor describing the compensation effect. This hypothesis is further tested in the second experiment as well as in the first experiment, using non-speech pure tone stimuli replicating the loudness contours of speech stimuli.

Table 1. Speech tokens chosen in the first experiment. The underlined CVC sequences are the target portions. The left most column shows the temporal positions of targets in a word.

Target position	Roman transcription				
C1V1C2	bakugeki	gakureki	hanareru	nagedasu	sakasama
C2V2C3	hanahada	imasara	kasanaru	katameru	mikakeru
C3V3C4	hanahada	korogasu	rokugatsu	tachimachi	tamatama

II. EXPERIMENT 1
— speech stimuli —

In the first experiment, the acceptability for temporal modification of V, C, or both V and C within a four-mora word was measured to test the two hypotheses (*CV* and *loudness*) on the temporal compensation effect.

Method

Subjects. Five adult subjects with normal hearing participated in the first experiment. All of them were native speakers of Japanese.

Stimuli. Fifteen four-mora Japanese words were chosen as the stimuli (see Table 1). The underlined CVC sequences were the targets of the modification; the temporal position of the target vowels were chosen from the first three out of four morae. It was confirmed in previous studies [7, 8] that each of the chosen V durations is sufficiently natural.

One of the V segments in each of the sequences and either the preceding or succeeding C were lengthened or shortened 15 or 30 ms using the LMA cepstral analysis-synthesis technique[12]. The modifications comprising a pair had the same absolute duration and either the same or opposite directions of change. The target portions were carefully trimmed out so as to exclude the transient portions at both ends of the Vs and the on-and-after burst portions of plosives or affricates. In addition to the above "double modified" stimuli, we prepared stimuli with a modification on V alone or on C alone and with an intact duration for reference. In total, 435 word stimuli were prepared.[1]

Procedure. The subjects listened to each of the word stimuli and were asked to rate the acceptability of temporal modification using seven subjective categories; i.e., "quite acceptable" to "unacceptable". Each subject rated a stimulus eight times in total. The obtained responses were pooled over all subjects for each category, then each stimulus was mapped on a unidimensional psychometric scale in accordance with Torgerson's Law of Categorical Judgment[13]. The scaled value of each "modified" stimulus was, then, adjusted by subtracting the scaled value of its corresponding "intact" stimulus. Thus, the obtained value for each stimulus corresponded to the amount of loss of acceptability from the intact one (reference).

Results and discussion

Figure 1 shows the loss of acceptability pooled over fifteen stimulus words for each of the four manners of temporal modification, i.e. V alone, C alone, V and C in opposite directions (V&C-opposite), or V and C in the same direction (V&C-same).

Let's assume the case of "double modification". If the two modifications are to perform the acceptability evaluation independently of each other, then the expected value for the loss of acceptability should be a simple sum of those values expected in the corresponding "single modification" cases. That is, the loss of acceptability in both the "V&C-same" and "V&C-opposite"

[1] They were, 15 CVCs x 29 variations of modification; i.e. 2 absolute modifications (= 15ms, 30ms) x 2 modification directions (lengthening, shortening) x 7 modification manners (= V alone, pre-C alone, post-C alone, CVsame, VC same, CV opposite, VC opposite) + 1 (= intact).

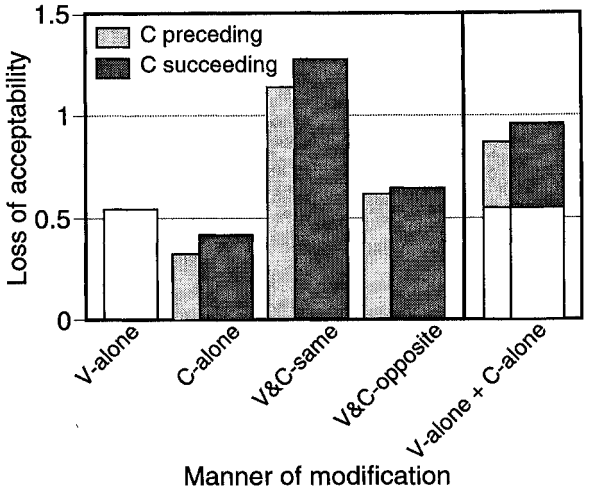

Fig. 1. Scaled loss of acceptability pooled over the fifteen word stimuli for each manner of temporal modification as a function of temporal order of V and C.

cases should approach the sum of those in the "V alone" and "C alone" cases (i.e., V-alone + C-alone). However, the value in "V&C-opposite" was smaller than that in "V-alone + C-alone" and the value in "V&C-same" was larger than that in "V-alone + C-alone" as shown in Figure 1. Multiple comparisons using Tukey-Kramer's HSD indicated both differences to be significant ($p < 0.05$).

These results mean that, (1) simultaneous modifications in a V duration and its adjacent C duration do not independently perform the acceptability evaluation, but that either (2) they perceptually compensate each other when in opposite directions, or (3) they perceptually enhance each other when in the same direction. This suggests that a unit having a time span larger than a single segment (C or V) functions in the time perception of speech.

To estimate the factors the perceptually dominant markers are related to, the following analyses focused on the modification manner of "V&C-opposite". If a large loss of acceptability were observed, this meant that some perceptually dominant temporal markers were located around the boundary of modified V and C. This is because modification in this manner does not suffer outside of the target V and C, but mainly moves the portion between them.

First, *the CV hypothesis* was evaluated. As shown in Figure 1, only a small difference in the loss of acceptability could be observed due to the temporal order of V and C. A *t*-test did not indicate this difference to be significant [$t(118) = 0.188$, $p = 0.851$]. Consequently, *the CV hypothesis* was not supported here. This result suggests that perceptually dominant markers do not generally locate around V-to-C boundaries.

Secondly, *the loudness hypothesis* was evaluated. Figure 2 shows two stimulus time-waveforms and their corresponding loudness contour which was calculated in accordance with ISO 532B[14] every 2.5 ms. As predicted from the examples of Figure 2, any of the V targets in this experiment was louder than its adjacent C portions; that is, each boundary between a modification pair always had a certain amount of loudness jump. In light of this fact, we adopted the loudness jump, calculated by subtracting the median loudness of C from that of V, as our explanatory variable in *the loudness hypothesis*. In addition to the factor of loudness jump, we included three factors capable of affecting the acceptability: the temporal order of V and C (V-to-C or

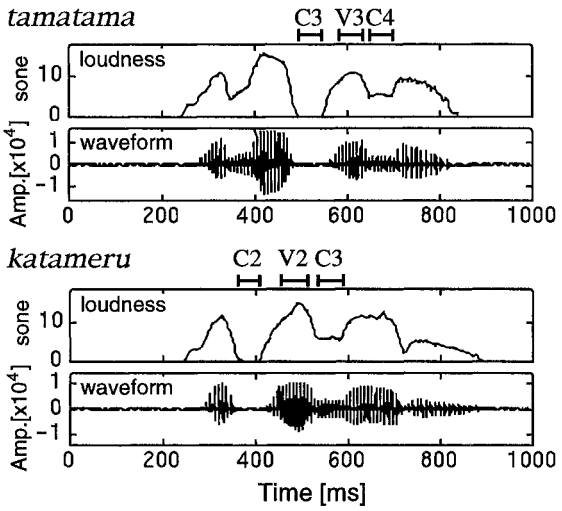

Fig. 2. Time waveforms and loudness contours of the word stimuli used in the first experiment. The horizontal bars on the top of each figure indicate the target portions to be modified. Upper: the word /tamatama/. Lower: the word /katameru/.

C-to-V), the temporal position of the modified vowel (1,2, or 3), and the amount of each single modification (15 ms or 30 ms).

The effects of the four factors mentioned above and their interactions, on the amount of loss of acceptability were tested by a four-way factorial General Linear Model (GLM)[15]. The main effect of the loudness jump was significant [$F(1, 96) = 10.51$, $p < 0.005$]. The loss of acceptability increased with increasing loudness jump. The interaction between the loudness jump and the amount of modification was significant [$F(1, 96) = 4.15$, $p < 0.05$]. The effect of loudness jump was higher for the longer (30 ms) modification condition. The temporal order of V and C was not significant again [$F(1, 96) = 0.087$, $p = 0.769$]. No other main factor nor interaction was significant.

There was no evidence for *the CV hypothesis* within the scope of the first experiment. On the contrary, the results of the GLM analysis supported *the loudness hypothesis*; a temporal displacement of a boundary with a large amount of loudness jump generally caused a considerable loss of acceptability. This suggests that perceptually dominant temporal markers tend to locate around major loudness jumps. Note, however, that there are several factors that were not included as the main factors that may affect the perception of temporal aspects of speech; e.g., a temporal dicriminability is higher around the phoneme boundary where the phonemic distinction depends on the durational cue [2]. Although we chose the stimulus tokens of the first experiment so as to balance such factors, they were not completely factored out. The second experiment was therefore designed to test whether the factor of loudness jump really affected the time perception, using non-speech stimuli replicating the time-loudness features found in the speech stimuli.

III. EXPERIMENT 2 — non-speech stimuli —

The purpose of the second experiment was to test the effect of loudness jump on the perceptual sensitivity to the temporal change of marker under controlled experimental conditions.

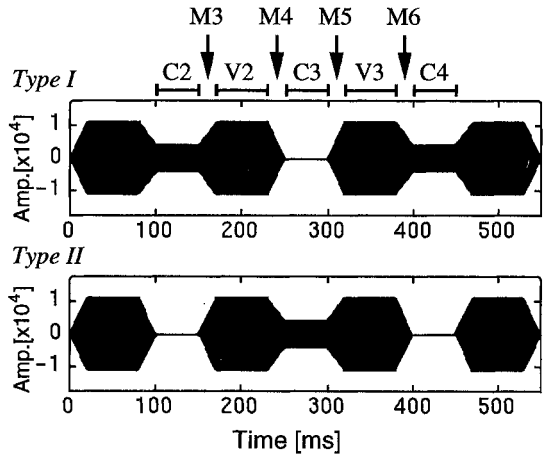

Fig. 3. Time waveforms of the two types of stimuli used in the second experiment. Each V or C indicates a target portion to be modified. Each M indicates the location of a considered temporal marker. The level of V is 73 dB SPL (= 9.77 sone), the level of louder C is 64 dB SPL (= 5.25 sone), and that of softer C is "silence". The duration of each slope is 10 ms (steep slope) or 20 ms (broad slope). All signals are 1 kHz pure tones.

Method

Subject. Six adult subjects with normal hearing participated in the second experiment.

Design. The experiment was designed as a four-way factorial one. The first factor was the loudness jump between two modified segments (large jump or small jump). The other three factors were included mainly to test their interactions with the first factor; they were, the direction of the marker slope (rising or falling), the steepness of the marker slope (steep or broad), and the temporal position of the marker in a sequence (2nd or 3rd).

Stimuli. Each stimulus was a 1-kHz pure tone with one of two types of amplitude contours as shown in Figure 3. These two types (type I and type II) were modeled on typical loudness contours of four-mora word stimuli (see Figure 2) and enabled us to complete a factorial design. Each stimulus comprised of the alternation of slope and steady portions; thus, only the slope portions could be the temporal markers. The steady portions each had one of the following three levels: 73 dB SPL, 64 dB SPL, or silence. On the loudness scale, they were 9.77, 5.25, and 0 sone, based on typical loudness values of vowels, nasals, and pre-burst closures, respectively. The duration of the slope was 10 ms (steep) or 20 ms (broad).

The duration of the loud portion (V portion) including rise-fall slopes and the soft or silent portion (C portion) were 100 and 50 ms, for the standard stimuli. One of the V durations in each of the comparison stimuli and either its preceding or succeeding C duration were modified in the opposite direction with 30 ms for each. The modification target was limited to the steady portions in the second V (V2) or the third V (V3) and either of its adjacent Cs (see Figure 3). Thus, the marker between the modified pair was solely displaced forward or backward by 30 ms from the standard. In total, 32 stimuli were prepared; they were: 2 types of amplitude contours (= type I, type II) x 2 steepness conditions (= steep, broad) x 2 slope directions (= rising, falling) x 2 target positions (= 2nd, 3rd) x 2 displacement directions (= forward, backward).

Procedure. In each trial, the subjects listened to the presentation of four successive stimuli, the first three each being the standard and the last one being a comparison, and they were asked to rate the difference between the standard and the comparison using eight numerical categories: "0" to "7"; the larger number corresponding to a larger subjective difference. Each subject participated in a one-hour preliminary training session. Twelve judgments were collected from

each subject for each stimulus. The obtained responses were pooled over all subjects for each category, then the detectability index, *d'*, for each comparison stimulus was estimated in accordance with the Theory of Signal Detection[16].

Results and discussion

A four-way completely randomized factorial analysis of variance (ANOVA) was performed for the obtained detectability *d'*. The factor of loudness jump was significant [$F(1, 16) = 99.5$, $p < 0.0001$]. The other three factors also turned out to be significant; they were, the direction of the slope [$F(1, 16) = 52.2$, $p < 0.0001$], the steepness of the slope [$F(1, 16) = 6.30$, $p < 0.05$], and the temporal position [$F(1, 16) = 36.7$, $p < 0.0001$]. Besides these main effects, a significant interaction was observed between the factors of loudness jump and slope direction [$F(1, 16) = 7.88$, $p < 0.05$]. No other interaction was significant.

Figure 4 shows *d'* for each marker condition pooled over the target positions and the marker displacement directions as a function of the loudness jump. As clearly shown in the figure, the effect of loudness jump agrees with the observed one in the first experiment; i.e., a larger loudness jump causes a higher sensitivity. This result supports *the loudness hypothesis* proposed in the first experiment.

The detectability of rising markers was significantly higher than that of falling markers. Therefore we thought that if we applied this effect directly to the first experiment, the displacements of the C-to-V transition (rising) would have a greater effect on the perception than the displacements of the V-to-C transition (falling). However, this was not the case. Further studies are necessary to understand the inconsistency between the factors of marker direction and temporal order of V and C.

The effect of temporal position in a sequence was significant; i.e., detection of the marker displacement around the second V was easier than that around the third V. This effect is consistent with Tanaka et al.'s finding[17] that the temporal discrimination for the initial interval is easier than that for the succeeding intervals in a click sequence.

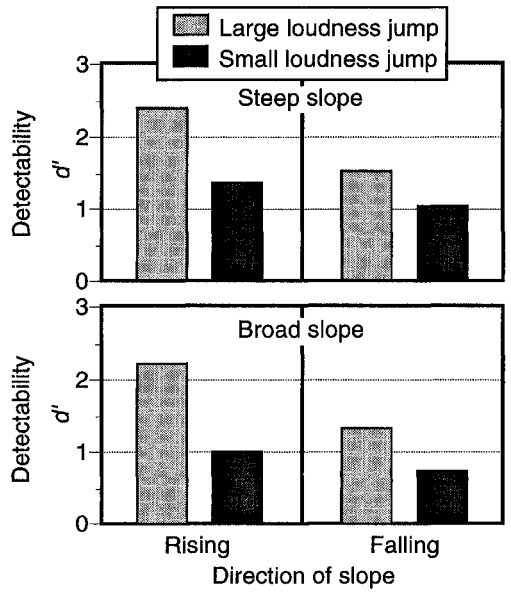

Fig. 4. Detectability index d' for a 30-ms displacement of a temporal marker in each stimulus condition, as a function of loudness jump between both sides of the marker. A larger d' implies easier detection.

IV. CONCLUSION

In the first place, it was confirmed that a unit having a larger time span, a moraic range or wider, than a single segment (C or V) seems to function in the time perception of speech; the perceptual compensation effect was generally observed between V durations and their adjacent C durations. Furthermore, it was suggested that the acoustical feature (loudness jump) is a more essential variable than the phonological or phonetical feature (CV or VC) to explain the perceptual compensation effect at such a higher unit. Large jumps in loudness were found to function as dominant temporal markers. Such large jumps generally coincide with the C-to-V and V-to-C transitions. This is probably one reason why previous studies have been successful, to some extent, in explaining perceptual phenomena by assuming a unit comprising CV or VC. However, the results of the current experiments indicated that the perceptual estimation is more highly correlated with the loudness jumps than with the level of CV or VC.

REFERENCES

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

[2] Fujisaki, H., Nakamura, K., and Imoto, T., "Auditory perception of duration of speech and non-speech stimuli," in Auditory Analysis and Perception of Speech, Fant, G. and Tatham, M. (Eds), Academic Press, London, pp. 197-219, 1975.

[3] Hoshino, M. and Fujisaki, H., "A study on perception of changes in segmental durations," (in Japanese with English abstract and English figure captions), Tech. Rep. H83-8, Acoust. Soc. Jpn., 1983.

[4] Sagisaka, Y. and Tohkura, Y., "Phoneme duration control for speech synthesis by rule," (in Japanese with English figure captions), Trans. Inst. Electron. Inf. Commun. Eng. Jpn, J67-A, pp. 629-636, 1984.

[5] Campbell, W.N. and Sagisaka, Y., "Moraic and syllable-level effects on speech timing," Tech. Rep. SP90-107, Acoust. Soc. Jpn., 1991.

[6] Schulze, H.-H., "The detectability of local and global displacements in regular rhythmic patterns," Psychol. Res., 40, pp. 173-181, 1978.

[7] Sato, H., "Segmental duration and timing location in speech," (in Japanese with English abstract), Tech. Rep. S77-31, Acoust. Soc. Jpn., 1977.

[8] Sato, H., "Some properties of phoneme duration in Japanese nonsense words," (in Japanese with English figure captions), Proc. Fall Meeting, Acoust. Soc. Jpn., pp. 43-44, 1977.

[9] Kato, H., Tsuzaki, M., and Sagisaka, Y., "Acceptability and discrimination threshold for distortion of segmental duration in Japanese words," Proc. ICSLP-92, pp. 507-510, 1992.

[10] Kato, H., Tsuzaki, M., and Sagisaka, Y., "Acceptability for durational modification of segments in words," (in Japanese with English abstract and English figure captions), Tech. Rep. SP92-145, Acoust. Soc. Jpn., 1993.

[11] Kato, H. and Tsuzaki, M., "Intensity effect on discrimination of auditory duration flanked by preceding and succeeding tones," J. Acoust. Soc. Jpn. (E), 15(5), 1994. (in press)

[12] Imai, S. and Kitamura, T., "Speech analysis synthesis system using the log magnitude approximation filter," (in Japanese with English figure captions), Trans. Inst. Electron. Commun. Eng. Jpn., J61-A, pp. 527-534, 1978.

[13] Torgerson, W.S., *Theory and Methods of Scaling*. New York: J. Wiley, 1958.

[14] ISO 532, "Acoustics - Method for calculating loudness level," International Organization for Standardization, 1975.

[15] SAS Institute Inc., SAS/STAT User's Guide. Vol. 2, Section 24 "The GLM Procedure", 1990.

[16] Green, D.M. and Swets, J.A., *Signal Detection Theory and Psychophysics*. New York: J. Wiley, 1966.

[17] Tanaka, M., Tsuzaki, M., and Kato, H., "Discrimination of empty duration in the click sequence simulating a mora structure," J. Acoust. Soc. Jpn. (E), 15, pp. 191-192, 1994.