Modeling and perception of temporal characteristics in speech

Yoshinori Sagisaka
GITI Waseda university, Tokyo and ATR, Kyoto

E-mail: sagisaka@giti.waseda.ac.jp, yoshinori.sagisaka@atr.co.jp

ABSTRACT

This paper describes characteristics of segmental duration control and its computational modeling that we have studied for more than two decades in speech synthesis. These studies not only contribute to prosody control in speech synthesis technology but also give an integrated view of individual temporal characteristics that have been found in phonetic science. The computational model can provide a new tool for analysis by synthesis of temporal characteristics by its prediction capability of assigning segmental duration in unseen contexts. Furthermore, a series of experimental results are shown on perceptual characteristics of duration modifications. These perceptual experiments reveal the context dependency of sensitivity to duration errors and strong correlation between duration errors and loudness that suggests the existence of a language universal temporal perception mechanism.

1. INTRODUCTION

Temporal control characteristics of speech have been studied for a long time for many purposes in different speech-related fields. Segmental duration characteristics have been measured over many languages to understand language universal/specific prosodic control by many phonetic scientists. In speech technology, fine control of segmental duration has been pursued to synthesize speech with natural rhythm and tempo. From human modeling viewpoints, cognitive scientists have tested production and perception hypotheses in order to better understand the human mechanisms of temporal control.

Throughout these research efforts in different fields, many control characteristics have been analyzed and common control factors and principles have been found. However, not all findings have yet been fully shared or integrated into one comprehensive model that can be used by speech researchers from all the different disciplines. Even by restricting to each academic field, it is sometimes quite difficult to understand the whole domain of control characteristics, underlying principles and mechanisms.

This paper sumarises our findings on segmental duration control for speech synthesis from more than a quarter of a century of research. Although all studies have been conducted for the primarily engineering purposes of synthesizing speech with natural prosody, many of the results can also be viewed as interesting scientific findings from other fields. By introducing a series of our research efforts in multiple research areas including acoustic phonetic analyses, computational modeling and perceptual characteristics analyses, we would like to introduce our current view on temporal control modeling and show how far we are now able to simulate human temporal processing mechanism in speech generation and perception.

In the following section, we present a series of analysis results on Japanese segmental duration characteristics. From a phonetic-science viewpoint, the dominance of a mora constraint as a timing unit is confirmed from multiple observations. In Section 3, computational modeling is proposed using statistical optimization. This model is not restricted to use in speech synthesis technology, but can also serve as a powerful tool for analyzing further characteristics. Finally, in Section 4, a series of subjective listening experimental results are described, using speech with temporal distortions. These perceptual investigations are not only useful for the quantitative evaluation of computational models but have revealed the high correlation between temporal characteristics and loudness.

2. FACTORS AND PRINCIPLES FOR THE CONTROL OF SEGMENTAL DURATION

2.1 The mora as a basic timing unit

In the temporal control of speech, there have been many studies aiming at simple control principles based on direct correspondences between segmental duration and phonetic notions. In the history of the ICPhS, there have been hot discussions where simple phonetically defined values (e.g. C/CV (the ratio between a syllable length and the length of its constituent consonant)) have been proposed as basic rhythmic units for western languages. For Japanese, moraic isochrony has been considered as a timing constraint and quite a few researchers have tried to find exact linear correlations between duration, length, and mora counts of utterances. As the analyses went on, it was confirmed that isochrony does not hold exactly in segmental durations [1]. Though some reasoning has successfully explained this mismatch from the discrepancies between production control and perception characteristics [2], exact calculation of utterance duration by phonetic unit counts appears too simplistic a control model.

We have measured segmental duration of Japanese using controlled, multiple, large speech databases and confirmed the existence of many control factors [3]-[7]. As shown in Table 1, these control factors range from local phonetic unit level to global sentence level. The difference arising from

ISBN 1-876346-48-5

the vowel and consonant categories appear to be dominant in average phoneme duration control. Vowels, /i/ and /u/are inherently shorter than /a/, /e/, and /o/. For consonants, unvoiced sounds tend to be longer than voiced ones. Fricatives tend to be longer than an plosives.

Segmental duration is not characterized only by these phonemic attributes but also by constraints from longer units. In particular, a moraic constraint is clearly observed in the control of Japanese vowel durations. A negative correlation is found between vowel durations and adjacent consonant durations. The temporal compensation of vowel duration is more influenced by the preceding consonant duration than the following one, and this is considered to be an acoustic manifestation of mora-timing. This temporal compensation has been observed both in raw duration data and in normalized z-score data for each vowel [8]. Through such statistical analyses, it has been confirmed that the compensation takes place in mora units but not in syllable units. Mora-timed rhythm of Japanese is in contrast to the stress-timed rhythm observed for other languages such as English. It has also been observed that not only segmental durations but pause length is under moraic control [9].

2.2 Phrase as a tempo unit for timing preset

There are also much longer control domains than the mora. At the phrase level, segmental durations change much more than in moraic units. The most remarkable phrase-level temporal control of segmental duration is local reset of speech tempo. Generally, the higher the mora count of a phrase, the shorter the average mora durations are in that phrase. Deviation from average mora length is greater in short pharases than in long phrases because of the higher freedom in short phrases. Long pharases cannot be produced at a slow tempo due to breathing constraints. The local tempo seems to be decided simply by phrasing. Moraic regularities are maintained within each phrase level unit. In each phrasal unit, the tempo is kept constant and there is no more local speech rate change.

The above mentioned moraic constraints and local phrasal tempo reset determines most of the variation in segmental durations of Japanese read speech. Though duration differences between content words and function words are well-known for English, only very small differences are found for Japanese. They are so small that only precise statistical analysis could have revealed the existence of differences [5]. Throughout these fine analyses, we have observed consistent lengthening in content words and shortening in function words. The lengthening characteristics look very coherent in a sense that semantically marked words such as a quantifiers and proper nouns have a larger degree of lengthening. Moreover, function words marking phrase structure, such as coordinate phrases or topicalized phrase boundaries, are often lengthened despite the usual shortening of other particles and auxiliaries. However, as already mentioned, these duration changes are typically less than ten milliseconds, and are almost always hidden by the bigger differences caused from moraic compensation and local phrasal tempo changes of more than a few tens of milliseconds.

2.3 Characteristics at tempo unit boundaries

As moraic constraints and phrasal timing resets determine the main temporal structure of Japanese read speech, there is not much freedom left in temporal control as seen from the small degree of word level differences. Among these, only changes at tempo unit boundaries are the exceptional. Shortning of phrase initial mora and lengthening of phrase final mora is quite marked. Remarkable shortening of sentence final mora has also been observed in declarative sentences. However, these temporal controls are restricted to only the single mora at the phrase boundary, which is in contrast to the continuous power decrease over multiple moras at phrase endings [10]. Unit boundary characteristics are observed only in phrases and sentences but not in words. Statistical analyses have shown that the word units themselves take part neither in boundary control nor in the decision of local tempo.

Table 1 Control factors of Japanese segmental durations

Range	Observed acoustic manifestations	Factor
Current phoneme	Intrinsic durations with very different deviations	Constraints in production
Neighboring phonemes Mora	Temporal compensation of neiboring phonemes Bi-moraic rhythm	Mora timing
Word	Content word lengthening Function word shortening	Markedness
Phrase endings	Moraic phrase final lengthening & initial shortening	Boundary making
Phrase	Uniform shortening inversely propotional to phrase mora counts	Local phrase tempo preset
Sentence	Total utterance length	Overall tempo

ISBN 1-876346-48-5

3. COMPUTATIONAL MODELING OF SEGMENTAL DURATION

3.1 Hierarchies and constraints in duration modeling

As seen in the previous sections, duration characteristics seem to be derived from multiple factors that form a control hierarchy of dominance. For Japanese, the moraic constraint and local phrasal tempo reset have the highest priority of control. Other factors can work only under these constraints. In computational modeling, these control hierarchies have already been implemented [1][9]. For Japanese [1], we first use an average duration for each consonant. Vowel duration is determined by adjusting the average vowel duration according to the length of preceding and following consonants, to take the moraic constraint into consideration. The consonant and vowel durations are then adjusted using mora counts in each phrase according to a local phrasal tempo. Finally vowel color differences and modifications at tempo unit boundaries are adjusted.

For British English too, hierarchical control has been employed by splitting the duration assignment into the syllable level and its constituent component levels [11]. Though there is very little similarity between our model for Japanese [1] and the early duration control model proposed by Denis Klatt for American English [12], there are many similarities to Campbell's model. These similarities suggest that the so called distinction between a mora-timed rhythm and a stress-timed one can be systematically and quantitatively explained by the difference in temporal units composing the mora or stress-group.

3.2 Corpus-based modeling

Although traditional rule-based control models assign reasonable values in most cases, serious errors are sometimes seen. These errors are often caused simply by the application of independently derived rules at the same time [13]. As large speech databases have become available, they have been used for modelling to prevent this type of error and to assign more accurate durations by applying statistical procedures for the prediction of segmental duration. As this corpus-based modeling approach has been quite successful, it is now widely employed in speech unit selection control for speech synthesis [14]. There are two big advantages to statistical modeling.

The first advantage is the accurate and fine modeling by itself. By introducing statistical optimization, there are no longer any large errors caused by the unexpected bad combinations of control rules. Furthermore, statistical techniques make it possible to analyse the very small but significant differences such as those between content words and function words in Japanese, as described in the previous section. The reduction of big errors definitely improves the naturalness of synthesized speech and the possibilities of fine analyses can provide us a good picture of a cognitive control model in phonetic sciences.

The second advantage is in the scientific basis underlying the corpus-based modeling. In conventional rule-based synthesis, as there is no clear description of data, control algorithms and error measures for modeling in many cases, it is quite difficult to trace and improve prediction systems. Whereas corpus-based modeling requires clarification of these three elements, everybody can share the knowledge and make use of the same control model. By carring out duration control experiments using a new test copus, we can find the limits of control accuracy and will be able to find ways to improve it by changing either the corpus, the control algorithms or the error measures. Thus we have attained a scientific system-building method to provide feedback error analysis results as a replacement for the empirical rule-based approach. It is expected that this corpus-based approach will become more commonly used in the phonetic sciences where each theory is usually tested under different conditions and with different data and measurements.

3.3 Statistical optimization for a duration model

For the control of duration, we have used the following linear regression model [3]-[7].

$$DUR = \mu(/*/) + \sum_{f} \sum_{c} Xfc \, \delta fc$$

In this equation, $\mu(/*/)$ denotes the mean duration of the current phoneme /*/, Xfc corresponds to the contribution coefficient of each category c of control factor f and δfc stands for the characteristic function of category c (i.e. δfc is 1 iff the current context corresponds to category c of f, otherwise 0) The control factors { f } correspond to e.g. current and neighbouring phoneme categories, mora counts of the phrase containing the current phoneme, and the current phoneme position, whose contribution is confirmed through statistical analyses, as shown in Table 1. In this formulation, {Xfc} values are pre-determined by the linear regression through a minimization of prediction error Σ(OBSERVED_DUR - DUR)^2 using a contextually balanced data set. This minimization is carried out simply by solving a normal equation which is gained by partially differentiating this error function as is standard in linear regressive analysis.

The duration prediction experiments using the speech data of 500 Japanese sentences showed that the root mean square errors were about 15 ms for both vowels and consonants [5][6]. Tests with both open and closed data showed error values that were comparable. It has also been quantitatively confirmed that this control can be applied to speech with different speaking rates or different speaking styles [7].

Additive-multiplicative modeling has been proposed as an extension of traditional linear analysis techniques for English, using bilinear expressions and statistical correlation analyses [15]. Alternatively, as the linear regression model has strong constraints in its functional form, regression trees have also been widely used [16]. In this model, no specific functional form is assumed for the prediction. A binary decision tree is formed by splitting the data set by specifying the contextual attributes of the segments. This method has been particularly applied for English duration control, where a larger number of segments and contextual differences need to be modeled.

To effectively reduce the control freedom in regression

ISBN 1-876346-48-5

trees by partially imposing constraints of linear models, Constrained Tree Regression(CTR) has been proposed [17]. In the CTR model, a superset of the traditional models, a regression tree is generated by controlling the tiedness of control factor parameters. By untying a shared parameter, or splitting one of the current leaves according to finer factor differences, more efficient use can be made of a new additional parameter freedom. By controlling the tiedness of the control parameters, CTR incorporates both linear and tree regressions as special cases and interpolates between them.

Though these statistical techniques can optimize duration control without losing freedom of conditioned exception control, they do not reflect control structure explicitly by themselves. Non-parametric statistical techniques show quite resonable performance if enough data is available. However, we need more clear temporal control architectures and mechanisms for much finer and efficient modeling of spech with various speaking styles.

3.3 Analysis by synthesis using a duration model

Until recently, duration models have been studied mainly based on read speech for text-to-speech applications. A duration model optimized using a read speech corpus has been applied to conversational speech [7]. Error analysis has shown that the segmental durations at phrase medial positions can be predicted fairly well in conversational speech. Big prediction errors have been observed at sentence final mora in particles indicating speakers' request for confirmation or agreement, strong conviction or assertion and interrogation.

As shown in this analysis, new control characteristics can be found by applying duration models to a new speech corpus with different speaking styles. Through error analysis, a computational model can serve as a good tool for *analysis by synthesis*. This synthesis-based analysis can give quantitative qualification of model correctness and supply good scientific evidence for cognitive reasoning.

3.4 Evaluation of a duration model

As seen in the previous sections, segmental duration is predicted in order to replicate the durations of natural speech in corpus-based speech synthesis. In this modeling, we commonly adopt the sum or average of errors between the calculated durations and the observed ones as the measure of evaluation. Though we can weight errors by modifying error functions, there is no theoretical reasoning in the lack of other knowledge.

It should be noted that the adoption of this error measure is based on assumptions that each distortion is contextually independent and that the subjective score is monotonously degraded by an increase in the sum of individual temporal distortions. These two implicit assumptions of the objective distortion criterion are not trivial (actually they are false). In section 4, we examine these two premises in the light of perceptual characteristics:

(1) A single duration distortion linearly correlates with the perceived distortion regardless of the attributes of the segment in question.

(2) Multiple duration distortions affect the perceived distortion independently of each other.

4. PERCEPTUAL CHARACTERISTICS OF SPEECH WITH TEMPORAL DISTORTIONS

4.1 Correlation between perceptual temporal distortion and loudness

Previous studies have reported that perceptual sensitivity to the duration change of a speech segment is affected by the segment quality; e.g., the temporal discrimination capability for vowel segments is higher than that of consonant segments [18]-[20], excluding the special case where a duration change also affects the phonemic category [21]. On the other hand, little is known about the acceptability of a change in segmental duration for the influence on segment quality, and this can be regarded as a more practical measure for the evaluation of duration control, although some other aspects of this measure have been explored in several studies [22],[23].

We have shown that listeners' rating scores of acceptability against changes in segmental duration can be accurately traced by a parabolic curve as shown in Figure 1 [24][25]. It has also been shown that the absolute value of the second-order coefficient of this approximation curve, namely the vulnerability index, is generally larger for vowel segments than for consonant segments (Figure 2, the left-hand scale). This tendency agrees with previous discrimination studies that vowel duration is more accurately discriminated than consonant duration.

Furthermore, it has been found that this variation in the vulnerability index is highly correlated with the intrinsic loudness of the segments as shown in Figure 2 (the right-hand scale). A non-speech study on temporal discrimination capability, on the other hand, showed that an auditory duration with large loudness is more accurately discriminated than a softer duration, if the target duration is temporally flanked by other sounds [26]. This tendency in temporal discrimination capability agrees with that of the acceptability measure found in Figure 2. All of these results suggest that the correlation observed between the vulnerability index (a sensitivity measure for acceptability) and the segment loudness can be accounted for as a reflection of the general characteristics of auditory perception. To take into account these perceptual characteristics, i.e., the dependency of duration sensitivity on segment quality, for distortion evaluation, the loudness characteristics should be added to approximate human subjective judgment more precisely.

4. 2 Perceptual interactions of temporal distortions in adjacent segments

A typical example of the interaction among multiple segmental distortions may be the perceptual compensation effect in the duration of two consecutive segments. This effect is like that when the duration of two segments is modified in a compensatory manner, i.e., to lengthen one segment and to shorten the other by the same amount; the total perceived distortion does not become very large in comparison with that expected from the sum of two independent modifications.

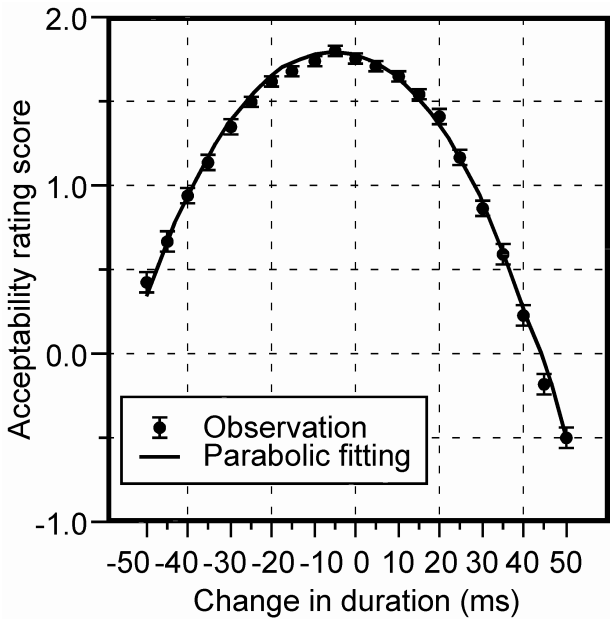

Figure 1 The change of subjective score in relation to duration modification
(An example of an acceptability rating profile as a function of change in the segmental duration. The dots and bars show the means and standard distortions of rating scores by six listeners using 70 segments in words. The parabolic fitting line is superimposed.)

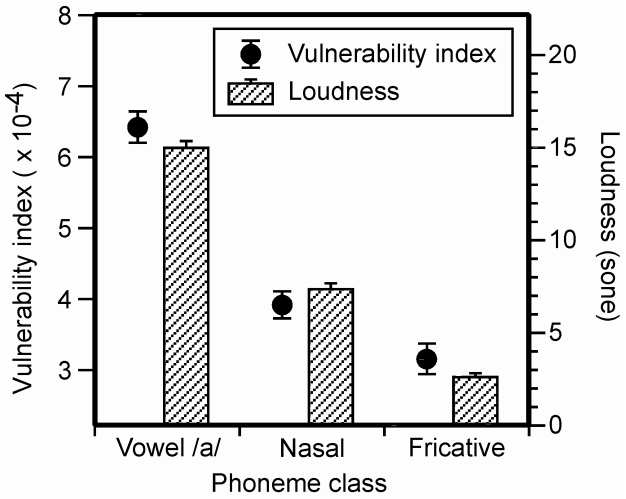

Figure 2 The change of subjective score in relation to duration modification
(The temporal vulnerability (the second-order coefficient of a parabolic fitting to acceptability rating scores with change in the segmental duration; dots, left-hand scale) and the loudness (bars, right-hand scale) of a speech segment as a function of phoneme class. The bars show the standard distortions. A larger vulnerability index implies a lower perceptual acceptability for a given change in the segmental duration.)

The perceptual compensation effect between consecutive vowel and consonant durations has been reported for both detectability of the modification [19][27] and acceptability rating [3][22][23]. The compensation effect of this sort indicates that the influence of a duration distortion is not limited within a segment but may extend beyond segmental boundaries, and also suggests that an evaluation criterion regarding each segmental distortion as independent is not perceptually valid.

Furthermore, it has been confirmed that the degree of perceptual compensation effect between two consecutive segments inversely correlates with the loudness difference or jump at the segmental boundary, in both detectability and acceptability tasks [28]. The amount of compensation decreased with increasing loudness.

A non-speech study also showed that the detectability of a compensatory temporal modification correlates with the loudness jump at the displaced boundary [28]. This suggests that the correlation observed between the perceptual compensation effect of speech and the loudness jump could be accounted for as a reflection of the general characteristics of the auditory perception.

Conventionally, while segmental distortions have been regarded as *changes in a segmental duration*, all of the above notions suggest that they can also be regarded as *the displacement of segmental boundaries*. For describing the relationship among multiple distortions, the latter view appears to be useful.

4. 3 Perceptually weighted error measure and further advances

We have combined the above perceptual characteristics into an evaluation measure [29]. In this measure we have taken into consideration two already known problems of conventional acoustic error measures. Using the simplifying loudness contour as an evaluation target parameter, we have successfully confirmed that our perceptual evaluation model with an objective measure can explain subjective listening scores better than the conventional acoustic error measures.

As we have focussed on the analysis of basic perceptual characteristics, mainly using isolated utterances, we have not yet captured the temporal perceptual characteristics of sentence speech. As a first step, we have conducted perceptual experiments on positional differences within a minor phrase [30]. As Figure 3 shows, we have found consistent positional effects. The temporal distortions are perceptually most sensitive at phrase initial positions and least sensitive at phrase final positions.

5. CONCLUSIONS

In this paper, I have surveyed segmental duration characteristics and computational models of temporal control for speech synthesis. Through the modeling processes, we have understood that many factors are related to timing control for speech generation, and that our perceptual mechanism of timing is derived systematically at very front level processing of loudness. From a speech engineering viewpoint, the corpus-based approach makes it possible to synthesize read speech with reasonable quality, and a computational model shows its usefulness for further modeling in the phonetic sciences. I am quite sure that the integration of all knowledge that we have in these different fields will bring us many more benefits in every field.

ISBN 1-876346-48-5

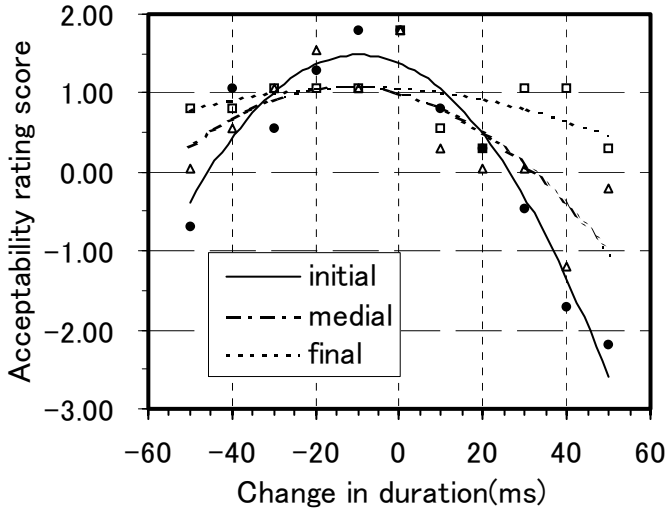

Figure 3 The difference of perceptual sensitivities at intra-phrase positions

(An example of one listener's perceptual sensitivities at phrase initial, medial and final position of /haruga/ embedded in sentence speech)

ACKNOWLEDGEMENTS

The author would like to offer particular thanks to many supervisors and collaborators. In particular, Yoh'ichi Tohkura, Hirokazu Sato and Shin'ichiro Hashimoto for their deep insights that have guided the underlying temporal control mechanisms. Kazuya Takeda, Nobuyoshi Kaiki, Hiroaki Kato, Naoto Iwahashi, Nick Campbell and Makiko Muto have supplied the author's knowledge continuously and worked to renew the control model step by step by their extensive endeavours. As easily understood from the references, most of works have been conducted by these people as principal investigators and the author is the luckiest person to have accumulated so much different knowledge directly from them at NTT, ATR, and Waseda Univ.

REFERENCES

[1] Kawasaki H.: "Models and data on the temporal regulation of speech : isochrony in Japanese and English" (in Japanese) JASJ Vol.39 No.6, pp.389-397, 1983

[2] Hiroya F. and Higuchi N.: "Temporal organization of segmental features in Japanese disyllables" JASJ (E) Vol.1 No.1, pp.25-30, 1980

[3] Sagisaka Y. and Tohkura Y.: "Phoneme duration control for speech synthesis by rule" (in Japanese) Trans. IElCE J67-A, No.7, pp.629-636, 1984

[4] Takeda K , Sagisaka Y and Kuwabara H.: "On sentential effects in the control of segmental duration in Japanese" JASA Vol.86 (6) pp.2081-2087, 1989

[5] Kaiki N., Takeda K and Sagisaka Y.: "Linguistic properties in the control of segmental duration for speech synthesis" p.255-264 in "Talking Machines" edited by G.Bailly et al North-Holland, 1992

[6] Kaiki N. and Sagisaka Y.: "The control of segmental duration in speech synthesis using statistical models" pp.391-402 in "Speech perception, production and linguistic structure" edited by Y. Tohkura et al Ohmsha IOS press, 1992

[7] Sagisaka Y.: "Prosody control for spontaneous speech synthesis" Proc. ICPhS pp.506-509, 1991

[8] Campbell N.: "A study of Japanese speech timing from the syllable perspective" (in Japanese) Trans. PSJ Vol.3 No.2,pp.29-39, 1999

[9] Kaiki N. and Sagisaka Y.: "Pause characteristics and local phrase-dependency structure in Japanese" Proc.ICSLP92 pp.357-360, 1992

[10] Mimura K., Kaiki N. and Sagisaka Y.: "Statistically derived rules for amplitude and duration control in Japanese speech synthesis" Proc. Korea-Japan joint workshop on advanced technology of speech recognition and synthesis pp.151-156, 1991

[11] Campbell W. N.: "Syllable-based segmental duration", in "Talking machines" Elsevier Science Publishers B. V. North Holland, pp.211-224, 1992

[12] Klatt D.H.: "Synthesis by rule of segmental duration in English sentences" in "Frontiers of Speech Comm. Res." edited by B. Lindblom et al (Academic Press) pp.287-299, 1979

[13] van Santen J. and Olive J. P..: "The analysis of contextual effects on segmental duration" Comp. Sp. and Lang. Vol.4, pp.359-390, 1990

[14] Sagisaka Y, and Campbell N.: "Description, acquisition, and evaluation of rules and data for speech synthesis" (in Japanese) Proc. IEICE pp.2068-2076, 2000

[15] van Santen J.: "Contextual effects on vowel duration" Speech Communication Vol.11 pp.513-546, 1992

[16] Riley M.D.: "Tree-based modeling of segmental durations" p.265-274 in "Talking Machines" edited by G.Bailly et al North-Holland, 1992

[17] Iwahashi N. and Sagisaka Y.: "Statistical modeling of speech segment duration by constrainied tree regression" Trans. IEICE Vol.E83-D, pp.1550-1559, 2000

[18] Huggins A. W. F.: "Just noticeable differences for segment duration in natural speech" JASA, Vol.51 pp.1270-1278, 1972

[19] Carlson R. and Granström B.: "Perception of segmental duration" in "Structure and Process in Speech Perception", edited by A. Cohen and S. Nooteboom (Springer-Verlag), pp. 90-106,1975

[20] Bochner J. H., Snell K. B. and MacKenzie D. J.: "Duration discrimination of speech and tonal complex stimuli by normally hearing and hearing -impaired listeners" JASA Vol.84, pp.493-500, 1988

[21] Fujisaki H., Nakamura K. and Imoto T.: "Auditory perception of duration of speech and non-speech stimuli" in Auditory Analysis and Perception of Speech", edited by G. Fant et al. (Academic Press), pp. 197-219, 1975

[22] Sato, H.: "Segmental duration and timing location in speech" (in Japanese), ASJ Trans. Tech. Comm. Speech S77-31, 1-8 1977

[23] Hoshino M. and Fujisaki H.: " A study on perception of changes in segmental durations" (in Japanese), ASJ Trans. Tech. Comm. S82-75, 593-5991983

[24] Kato H., Tsuzaki M. and Sagisaka Y.: "Acceptability for temporal modification of single vowel segments in isolated words," JASA Vol. 104, pp.540-549, 1998

[25] Kato H., Tsuzaki M. and Sagisaka Y.: "Effects of phoneme class and duration on the acceptability of modifications in speech" JASA. Vol. 111, pp. 387-400, 2002

[26] Kato H. and Tsuzaki M.: "Intensity effect on discrimination of auditory duration flanked by preceding and succeeding tones" JASJ (E) Vol.15, pp.349-351, 1994

[27] Huggins A. W. F.: "On the perception of temporal phenomena in speech" JASA Vol. 51, pp.1279-1290, 1972

[28] Kato H., Tsuzaki M. and Sagisaka Y.: "Acceptability for temporal modification of consecutive segments in isolated words" JASA. Vol. 101, pp.2311-2322, 1997

[29] Kato H., Tsuzaki M. and Sagisaka Y.,: "A modeling of the objective evaluation of durational rules based on auditory perceptual characteristics" Proc. ICPhS pp.1835-1838, 1999

[30] Muto M., Kato H., Tsuzaki M., Sagisaka Y.: "Effect of intra-phrase position on acceptability of changes in segmental duration in sentence" Proc. ICSLP pp.761-764, 2002