COMPUTATIONAL MODELING OF TIMING CONTROL AND ITS APPLICATION TO OBJECTIVE EVALUATION OF THE SECOND LANGUAGE PROFICIENCY

Yoshinori Sagisaka[(1)], Hiroaki Kato[(2)], Minoru Tsuzaki[(3)] Shizuka Nakamura[(1)] and Chatchawarn Hansakunbuntheung[(4)]

[(1)] GITI / Language and Speech Science Research Lab., Waseda University
[(2)] NICT / ATR Media Information Science Laboratories, [(3)] Kyoto City University of Art,
[(4)] HLT Lab., National Electronics and Computer Technology Center

Abstract: In this article, we introduce the studies on computational modeling of segmental duration that we have been conducting for decades to find out speech timing control factors and mechanism. In addition, some experimental results are also shown on loudness dependence in duration perception. These duration modeling and perceptual characteristics on duration error sensitivity to loudness give us hints for objective evaluation of the second language (L2) proficiency. Some experimental results are introduced to show the usefulness of duration contrl model and the need for perceptual studies for objective L2 proficiency evaluation.

1 Introduction

By ceaseless efforts of many speech researchers, speech timing characteristics have been analyzed and modelled in many languages. Many control factors and principles have been found and modeled for each language and reasonable prosodic characteristics have been obtained to synthesize speech using statistical optimization schemes. However, all findings have not yet been fully shared by or integrated into single comprehensive model where language independent and dependent control characteristics can be systematically explained. In application fields such as speech synthesis, though we have obtained computational models with reasonable prediction precision, scientific understanding of language dependency in timing control and its perception have not yet been throughly studied.

Throughout our studies of the timing modeling and evaluation for decades, we noticed that various types of knowldge in science fields such as phonetics, psychology are needed to properly build speech information technology. In this paper, we will introduce our findings on computational modeling of segmental duration control and the perceptual characteristics on timimg. Through a series of our research in multiple research areas including acoustic phonetic analyses, computational modeling and perceptual characteristics analyses and an application to the second language proficiency evaluation, we would like to show the need of further studies on a language common timing modeling where we can compare timing control characteristics cross linguistically based on human processing mechanism in speech generation and perception.

In Section 2, we briefly introduce analysis results on segmental duration characteristics to show control factors and principles for Japanese. In Section 3, we exlain a computational segmental duration model using statistical optimization. The obtained duration model has been used to assign segmental durations in speech synthesis. In Section 4, some experimental results on perceptual characteristics are introduced for speech with temporal distortions. The perceptual study shows high correlation between timing characteristics and loudness, which provides a scientific reasoning to design an objective error measure reflecting perceptual characteristics. In Section 5, we show experimental results of objective evaluation on duration to asess second language speech learners' proficiency. Finally, we sum up our findings and

introduce a research effort of Asian English Speech Corpus Project (AESOP) towards cross linguistic study of speech uttered by the second language learners.

2 Control factors and priciples in timing

2.1 Moraic constraints in Japanese timing

In the timing control of speech, there have been many studies aiming at simple control principles based on direct correspondences between segmental duration and phonetic notions. 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]-[6]. As shown in Table 1, these control factors range from local phonetic unit level to global sentence level. The difference arising from the vowel and consonant categories appear to be dominant in average phoneme duration control. 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. Through 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 [7].

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 , moraic final shortning	Overall tempo

2.2 Local tempo preset for phrases

There exists much wider control range than mora. In Japanese timing, local timing preset is the most remarkable one. Generally speaking, 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.

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]. 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. Only changes at tempo unit boundaries are seen. At phrase final mora, lengthening is observed and remarkable shortning can be seen at sentence final mora [4].

3 Computational modeling of segmental duration

3.1 Duration modeling using statistical optimization

To assign phone durations for speech synthesis, a lot of duration models have been proposed. For English, to reduce prediction errors of original rule-based models such as Denis Klatt's model for American English [8], corpus-based models have been proposed (e.g. [9]-[12]). In these models, statistical optimization schemes have been widely used to minimize total prediction errors. For Japanese, we have used the following linear regression model [3]-[6].

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

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 neighboring 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 $\Sigma(\mathrm{OBSERVED\_DUR} - \mathrm{DUR})^2$ 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]. 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 [6]. To effectively reduce the control freedom in regression trees by partially imposing constraints of linear models, we have proposed Constrained Tree Regression(CTR) model [13]. 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.

3.2 Reconsideration on predicted duration errors

In the current duration modeling including our modelings, we employ statistical optimization to minimize prediction errors defined as root square means. Though this optimization measure looks quite normal from engineering viewpoint, we have to understand what the adoption of this error means. It should be noted that the adoption of this purely acoustic error measure is based on assumptions that each distortion is contextually independent and that the subjective speech quality 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 the following Section 4, we will introduce the analysis results on temporal distortions to 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

Though statistical optimization enables to reduce phone duration errors, the goal of duration model in speech synthesis application is to generate natural timing. To measure perceptual relevance, we need subjective listening score on naturalness. As a subjective measure, MOS (Mean Opinion Score) showed supperior sensitivity to other measures such as pair comparisons. We found that the relationship between duration distortions and MOS scores of resultant speech quality can be expressed by a parabolic curve and that the vulnerability index (i.e. absolute value of the second-order coefficient) of this approximation curve is generally larger for vowel segments than for consonant segments [14]-[15]. Furthermore, it has been found that the vulnerability index is highly correlated with the intrinsic loudness of the segments. 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 [16]. 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.

In addition to duration distortions in a single segment, there is a perceptual compensation of duration modifications between adjacent segments [3][17]. We can find 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 [17]. The amount of compensation decreased with increasing loudness. Perceptual studies have been kept on for other contextual differences such as positional differences within a minor phrase [18] and speech rate changes [19].

A non-speech study also showed that the detectability of a compensatory temporal modification correlates with the loudness jump at the displaced boundary [16]. 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.

4.2 Loudness weighted error measure for temporal distortions

We have integrated the above perceptual characteristics into an evaluation measure [20]. In this new 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.

Though this perceptually weighted distortion measure was proposed to evaluate speech quality degradation of synthetic speech due to timing irregularities, we can expect its effectiveness for other applications to evaluate temporal distortions subjectively. In the next section, we will show how the duration model and the loudness weighted error measure for temporal distortions can be used to evaluate the second language learner's proficiency.

5 Evaluation of L2 proficiency based on timing characteristics

5.1 Objective evaluation of L2 learner's timing characteristics

Since there is large language-dependency in timing control factors as the difference between stress-timing and syllable-timing, timing control is an important issue in second language (L2) learning. However, as same as perceptual characteristics on duration changes, there are not so many studies to measure L2 learner's prosody control proficiency objectively. For better L2 education and fundamental understanding of timing control mechanism, we have started to analyze learner's timing characteristics by comparing duration differences between English native speakers and non-native learners using identical English sentences [21]-[24].

Correlations between raw duration differences and subjective evaluation scores clearly showed the existence of both language-dependent factors and learner reading proficiency [21]. To obtain an objective measure which has higher correlation between MOS subjective evaluation score, we have analyzed the effect of sentence length and tempo normalization to raw duration differences. The correlation analyses between subjective evaluation scores showed high correlation in longer sentence sets and the effectiveness of tempo normalization [22]. These facts suggest the possibility of better prediction by looking for much finer measurements and measurement targets.

5.2 Model and measure for temporal distortions

Motivated by our findings throughout duration studies, we have been continuously adopting our knowledge on duration characteristics explained in the previous sections for a better evaluation model and an effective measure. First, we applied a duration model by itself to factor out all contextual effect [23]. As explained in Section 3, a statistical duration model employs quite many contextual variables. This means that a duration model can be employed to directly compare the duration differences between learners and natives by contextual

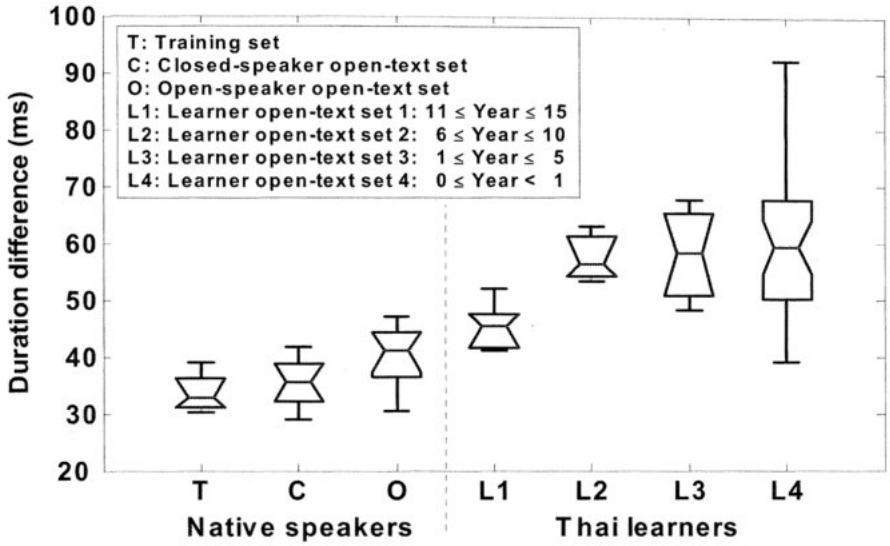

Figure 1 Comparison of RMS duration differences from predicted durations between English natives and Thai learners (C: native closed speakers, O: native open speakers, L1 – L4: Thai learners grouped by education period in English-as-an-official-language countries)

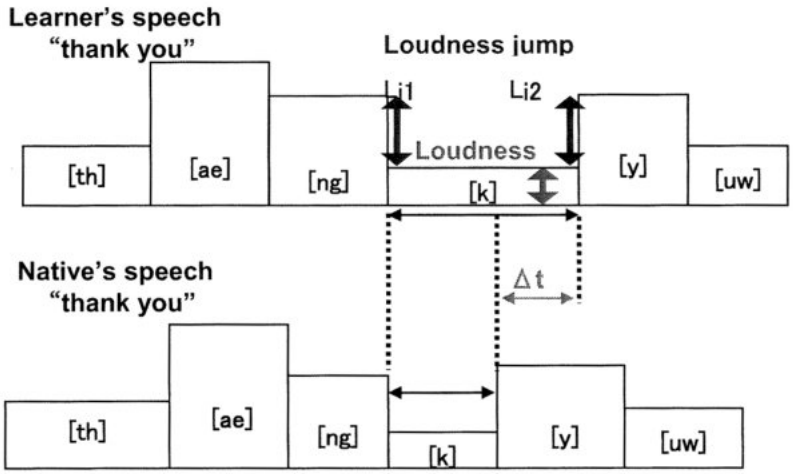

Figure 2 Duration difference weighting using the loudness of the current segment and the loudness jumps from adjacent segments (Duration difference Δt is weighted by w (Lc, Lj1, Lj2) using loudness of the current segment Lc and Loudness jump from adjacent segments Lj1 & Lj2)

effects. An application of English duration model trained using English native's data has clearly shown non-native learner's English proficiency. subtracting

Figure 1 shows the duration differences from predicted durations between English native speakers and Thai learners grouped by English-education experience in English-as-an-official-language countries [23]. As shown in this Figure, noticeable duration differences were

observed by learners' grouping according to the time spent in English education. The duration differences of the close speaker open-text set showed the least difference from that of the training set (i.e., the close-speaker close-text set). This group also showed the smallest duration differences among all speaker groups. Accordingly, the results showed consistency and reasonable prediction accuracy of the model both for the training and for the open set.

Not only the use of duration model, but also the perceptual weighting has turned out to be useful for an objective evaluation measure [24]. As shown in Figure 2, by the weighting of the raw duration differences using the loudness of the current segment and the loudness jump from adjacent segments, we can get a new objective measure reflecting perceptual characteristics. The correlation score 0.54 between the raw duration differences and subjective MOS scores increased to 0.72 by using this new loudness weighted difference measure, This remarkable correlation increase is quite promising to pursue a new objective measure by integrating our knowledge on segmental duration characteristics.

6 Conclusion

In this paper, we have introduced segmental duration characteristics and computational models of timing control. Through the modelling, we have shown 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. As an example of application of duration modelling to the objective evaluation of the second language speech, both computational models and measures have turned out to be useful.

To carry out the research introduced in this paper, we have been using a lot of speech and language corpora. We could have got none of these research results without them. To obtain full annotated rich data, it takes huge human efforts and time for content design, data collection, various types of transcriptions and preparation for associated information such as speaker's background and natives' subjective evaluation scores for L2 data. We will definitely need more data to expand our research area to neighboring related field of education or other basic science.

To carry out our research and enhance collaborations in different countries and multi-disciplinary research fields, in particular L2 related fields, we set up the consortium of AESOP (Asian English Speech cOrpus Project) where researchers in different countries have started to work together in 2008 fall. The first target of AESOP is to build up a common English speech corpus which represents the varieties of English spoken in Asia. We would like to form an international consortium of linguists, psychologists, speech scientists, technologists and educators to work on L2 speech and language. As the first step, we would like to collect and compare English speech corpora first from the Asian countries using a consistent set of core materials in order to derive a set of phonetic properties common to all varieties of Asian English [25] and provide a research platform that we can share. Definitely, we would like to expand our research collaboration to other countries and languages in near future to have better understanding of language common and language dependent phenomena cross linguistically. We are quite sure that the integration of all knowledge that we have in different fields will bring us many more benefits in every related field.

Acknowledgements

The authors would like to express many thanks to 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, Naoto Iwahashi, Nick Campbell and Makiko Muto have supplied our knowledge continuously and worked to renew the control model at NTT, ATR, NICT and Waseda University.

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. IEICE 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. 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

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

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

[8] 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

[9] Pitrelli J. and Zue V.: "A hierarchical model for phoneme duration in American English" Proc. European Conf. on Speech Communication Technology, 1990

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

[11] 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

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

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

[14] 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

[15] 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

[16] 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

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

[18] Muto M., Kato H., Tsuzaki M. and Sagisaka Y.: "Effect of intra-phrase position on acceptability of change in segment duration in sentence speech" Speech Communication 45 pp. 361-372 2005

[19] Muto M., Kato H., Tsuzaki M. and Sagisaka Y.: "Effect of speaking rate on the acceptability of change in segmental duration" Speech Communication 47 pp.277-289 2005

[20] 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

[21] Muto, M., Sagisaka Y., Naito T., Maeki D., Kondo A., Shirai K.: "Corpus-based modeling of naturalness estimation in timing control for non-native speech" Eurospeech pp.498-501, 2003

[22] Nakamura S., Tsubaki H., Kondo Y., Nakano M. and Sagisaka Y.: "Tempo-normalized measurement and test set dependency in objective evaluation of English learners' timing characteristics" Proc. 16th ICPS pp.1733-1736 2007

[23] Hansakunbuntheung, C., Kato, H., and, Sagisaka, Y.: "Model-based automatic evaluation of second-language learner's English segmental duration characteristics," Acoustical Science and Technology Vo.31 No.4 pp.267-277, 2010

[24] Nakamura S., Matsuda S. , Kato H., Tsuzaki M. and Sagisaka Y.: "Objective evaluation of English learners' timing control based on a measure reflecting perceptual characteristics", Proc. IEEE ICASSP, pp.4837-4840, 2009

[25] Visceglia T., Tseng C., Kondo M., Meng M. and Sagisaka Y.: "Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project) " Proc. O-COCOSDA CDROM Q2-1 pp.67-72 2009