OBJECTIVE EVALUATION OF ENGLISH LEARNERS' TIMING CONTROL BASED ON A MEASURE REFLECTING PERCEPTUAL CHARACTERISTICS

Shizuka Nakamura[1, 2], Shigeki Matsuda[3], Hiroaki Kato[4], Minoru Tsuzaki[5], Yoshinori Sagisaka[1, 2]

[1]GITI, Waseda University / [2]Language and Speech Science Research Laboratories
[3]NICT / ATR Spoken Language Communication Research Laboratories
[4]NICT / ATR Cognitive Information Science Laboratories
[5]Kyoto City University of Arts, Japan

ABSTRACT

Automatic evaluation of English timing control proficiency is carried out by comparing segmental duration differences between learners and reference native speakers. To obtain an objective measure matched to human subjective evaluation, we introduced a measure reflecting perceptual characteristics. The proposed measure evaluates duration differences weighted by the loudness of the corresponding speech segment and the differences or jumps in loudness from the two adjacent speech segments. Experiments showed that estimated scores using the new perception-based measure provided a correlation coefficient of 0.72 with subjective evaluation scores given by native English speakers on the basis of naturalness in timing control. This correlation turned out to be significantly higher than that of 0.54 obtained when using a simple duration difference measure.

Index Terms— Educational technology, Oral communication, Cognitive science, Psychology

1. INTRODUCTION

To automatically evaluate second-language learners' proficiency, many studies have aimed to capture the differences between English native speakers and learners. Most of them concentrated on segmental quality evaluation where statistical speech recognition frameworks have been employed to measure phonetic differences, e.g. [1-4]. Some studies have concentrated not only on segmental quality but also on prosody [5-8].

Since there is large language-dependency in timing control factors such as the difference between stress timing and syllable timing, timing control is one of the most crucial issues in the learning of second-language prosody. As a first step for objective evaluation of English learners' timing control, we have analyzed learners' timing characteristics from the viewpoint of duration by measuring duration differences between English native speakers and learners using identical English sentences [7-8].

The results obtained show that duration differences between English native speakers and learners correlate with subjective evaluation scores of learners' English timing control [7-8]. However, these correlations were in general not high enough for practical use. In these studies, it was also observed that the speech duration of learners tended to be longer than the corresponding duration by native speakers, i.e., learners spoke more slowly than native speakers [7].

Since duration differences caused by tempo differences extended over every speech unit, they may conceal other factors in analyses of simple duration differences. As one solution to this problem, we adopted tempo normalization to analyze duration differences. Tempo normalization by the phone-specific lengthening and shortening characteristics provided better results but the effect was limited; higher correlations were achieved only when using longer test sentences, which clearly include control tendency [7]. However, the characteristics found in evaluating learners' timing control are not necessarily explained solely by the durational aspect of learners' speech. The same physical quantity of duration distortion may cause different perceptual outcomes depending on the attribute and context of the target segment [11]. Such factors affecting perceptual evaluation include loudness-related ones.

In this paper, we adopted a new objective evaluation measure instead of the previous evaluation measure based on duration differences only. The new measure is taking perceptual characteristics into account. The duration difference is weighted by the loudness of the corresponding speech segment and the differences or jumps in loudness from the two adjacent speech segments. First, we introduce objective evaluation using the previous acoustical measure, based on duration differences, in the following section. Next, we explain the new measure reflecting perceptual characteristics in Section 3. Finally, we run objective evaluation experiments of English learners' timing control by using the measure reflecting perceptual characteristics in Section 4.

2. OBJECTIVE EVALUATION OF ENGLISH LEARNERS' TIMING CONTROL BASED ON DIFFERENCES IN ACOUSTICAL FEATURES

2.1. Objective evaluation based on duration differences

For objective evaluation of learners' timing control, we analyzed duration differences between native speakers and

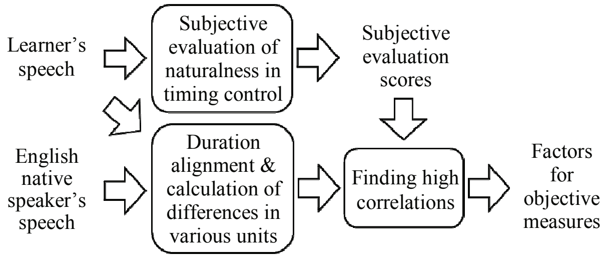

Learner's speech
Subjective evaluation of naturalness in timing control
Subjective evaluation scores
English native speaker's speech
Duration alignment & calculation of differences in various units
Finding high correlations
Factors for objective measures

Figure 1: *Procedure to find objective evaluation factors.*

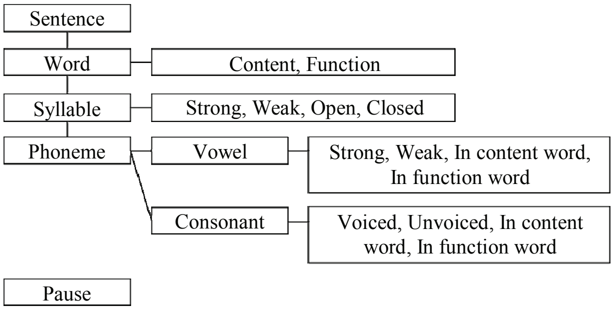

Sentence
Word
Content, Function
Syllable
Strong, Weak, Open, Closed
Phoneme
Vowel
Strong, Weak, In content word, In function word
Consonant
Voiced, Unvoiced, In content word, In function word
Pause

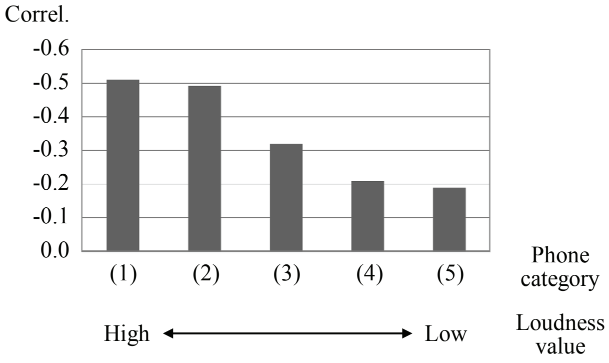
Figure 2: *Examples of phonetic speech units.*

learners as an objective measure. Figure 1 shows the procedure used to find objective evaluation factors. As shown in figure 1, we analyzed correlations of subjective evaluation scores of learners' timing control given by native English evaluators with duration differences between English native speakers and learners. Duration differences are calculated for each of the phonetic speech units listed in figure 2. As a result of these analyses, we found comparatively strong correlations of duration differences in the following speech units: sentence, weak syllable, pause, vowel in function word, and word [7]. By using these factors, we developed an objective evaluation model to predict subjective evaluation scores by adopting a linear multiple regression model.

2.2. Analyses on differences in contribution of evaluation factors to subjective evaluation based on perceptual characteristics

As a result of further analyses of evaluation factors to get stronger correlations with subjective evaluation scores, we obtained the result that differences in contribution of evaluation factors could be explained by perceptual characteristics. Figure 3 shows example results of correlation analyses of duration differences with subjective evaluation scores for the phone segment at the end position in each syllable. As shown in figure 3, vowel segments rather than consonant segments show stronger correlations; furthermore, stronger vowel segments rather than weaker vowel segments show stronger correlations. This tendency was consistently observed in most of the other data.

By carefully observing the correlation differences, we noticed that this result was in good agreement with a previous study showing that the durational error of a vowel segment is perceived as more salient than that of a consonant segment [11]. This suggests the usefulness of loudness-related weighting because a strong vowel is in general louder than a voiced/unvoiced consonant. Therefore, we examine the evaluation measure by giving consideration to loudness. In this paper, we use approximate values of loudness calculated according to the International Standard, ISO-532B, from given waveforms. We use the term 'loudness' with such a limited meaning in the following part. To make this paper's presentation comprehensive, we briefly explain the time-loudness marker model [9-10], which we use as a new evaluation measure, in the next section. Then we describe our evaluation experiments on the new measure in Section 4.



Strong vowel (1: primary, 2: secondary, 3: tertiary)
Consonant (4: unvoiced, 5: voiced)

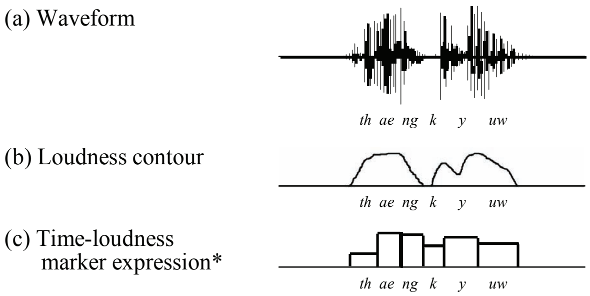
Figure 3: *Example results of correlation analyses of duration differences with subjective evaluation scores in phone segment at the end positions in each syllable.*

3. AN EVALUATION MEASURE OF TIMING CONTROL BASED ON PERCEPTUAL CHARACTERISTICS

3.1. Time-loudness marker model

The time-loudness marker model [9-10] has been proposed to evaluate naturalness in timing control. This model reflects perceptual characteristics by using a physical quantity based on loudness. The time-loudness marker model predicts human acceptability of duration distortion by using the duration and loudness of speech segments calculated from the acoustic features of their waveforms.

The time-loudness marker model simplifies the loudness information of a given speech sample by taking representative loudness values in each speech segment to obtain the time-loudness marker expression. Figure 4 shows the procedure of extracting a time-loudness marker expression (i.e., discrete loudness values). After calculating the loudness contour (b) from a given waveform (a), the representative loudness values are extracted in every phone to form the time-loudness marker expression (c). In this paper, the average loudness value is taken as the representative loudness for each phone.



(*Average loudness scores in every phone segment)

Figure 4: *The procedure of extracting time-loudness marker expression (i.e. discrete loudness values).*

The time-loudness marker model predicts the acceptability decrement against a given duration distortion. A greater acceptability decrement means more difficulty in perceiving speech with naturalness. Figure 5 shows example time markers for the test sentence 'Thank you.' The time-loudness marker model calculate the overall acceptability decrement of a given speech sequence, e.g., word or sentence, from its reference, e.g., a native-speaking sample, on the basis of the change in time interval between two perceptually salient time points or markers with loudness-related weightings. In this paper, we assume the markers coincide with the beginning and ending points of each phone. First, we calculate the perceptual weighting factor (W_{ij}) for the interval between i-th and j-th markers (t_{ij}) by using equation (1) with an invariable (b) for scale adjustment, the loudness jumps at the i-th and j-th markers (l_i and l_j, respectively), and the representative loudness of the speech section between the two markers (C_{ij}).

$$w_{ij} = b\frac{(l_i + l_j)}{2} + C_{ij} \tag{1}$$

Next, the acceptability decrement (l_{ij}) caused by the duration change (Δt) of the interval between the i-th and j-th markers is calculated by equation (2) with an invariable (a) for scale adjustment.

$$l_{ij}(\Delta t) \cong \frac{a \cdot w_{ij} \cdot \Delta t^2}{\sqrt{t_{ij}}} \tag{2}$$

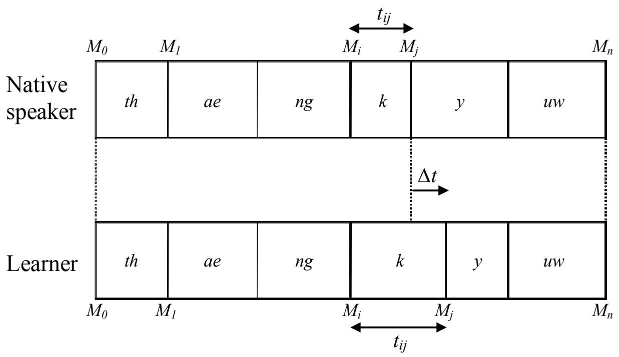
Finally, the summation of acceptability decrement (L) in a whole sentence to given duration distortion is calculated by equation (3).

$$L = \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} l_{ij} \tag{3}$$

In a previous study for evaluating the naturalness of synthesized speech [10-11], the effectiveness of the time-loudness marker model was verified in objective experiments of duration distortion.

3.2. Perceptually weighted evaluation measure for English learners' timing control

Because of its capability of naturalness evaluation, we expect that the time-loudness marker model is also effective for second-language learners' speech. In this paper, we run objective evaluation of English learners' timing control by using the time-loudness marker model. The calculation model and the application procedure followed a previous suggestion [10] by regarding the acceptability decrement as a value showing the unnaturalness of learners' speech.



M_i, M_j : i-th and j-th markers
t_{ij} : Interval between two markers M_i and M_j
Δt : Duration differences of t_{ij} between the native speaker's reference sample and learner's one

Figure 5: *Example time markers for the test sentence of 'Thank you.'*

Since the time-loudness marker model predicts the acceptability decrement against a given duration distortion, it is expected that the acceptability decrement has a certain correlation with the subjective evaluation scores. In the next section, we run objective evaluation experiments by using our new evaluation measure reflecting perceptual characteristics.

4. EXPERIMENTS ON EVALUATION OF ENGLISH LEARNERS' TIMING CONTROL USING A WEIGHTED MEASURE

We estimated objective evaluation scores of naturalness in timing control by using the time-loudness marker model. We adopted 231 learners' speech samples, which consisted of the same phone sequence as the corresponding reference by native speakers, from an English speech database read by Japanese students [12]. The speakers were 75 university students whose native language was Japanese. They had a wide range of reading proficiencies in English. To serve as reference samples, we selected speech samples spoken by two English language teachers who spoke General American.

To predict subjective evaluation scores of naturalness in timing control, a linear multiple regression model was adopted. To compare the prediction performance of different measurements and sentence sets, models were trained under six different conditions: the combinations of two weighting conditions (with or without time-loudness marker model (i.e., the latter is simple duration difference model)) and three sentence-set conditions (very long (VL), very short (VS), or all four degrees of sentence length (ALL), whose examples are shown in Table 1). ALL, VL, and VS sentence sets consisted of 231, 60 and 60 samples, respectively. For model training in

Table 1: *Example test sentences consisted of four degrees of sentence length.*

Sentence length	Test
Very Short (VS)	I'm amused.
Short (S)	I'm amused by the man.
Long (L)	I'm amused by the man and his jokes.
Very Long (VL)	I'm amused by the man and his very funny jokes.

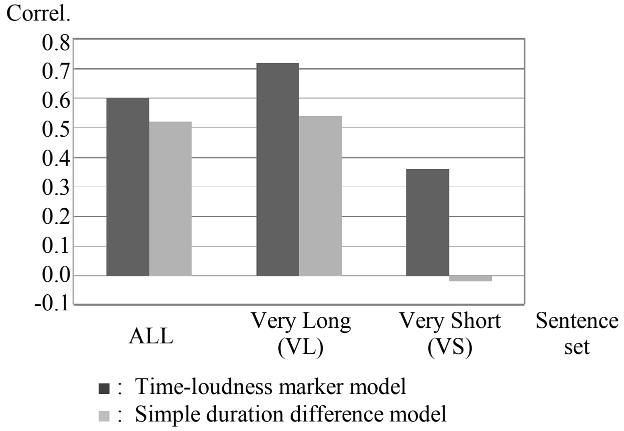


Figure 6: *Correlations between subjective evaluation scores and predicted scores under six differences conditions.*

each sentence set, two-thirds of the samples were used, and the rest of the samples were used for the test.

Figure 6 shows the correlations between subjective evaluation scores and predicted scores for open data. A higher correlation implies a better prediction, i.e., closer to human evaluators. Objective evaluation models using the time-loudness marker model achieved higher correlations than the simple duration difference model. It is noteworthy that a correlation of 0.37 was obtained for the VS sentence set when using the time-loudness marker model while no significant correlation was observed when using the simple duration difference model. As a result of the increase in correlation of approximately 0.2, the correlation for the VL sentence set using the time-loudness marker model was over 0.7.

5. CONCLUSIONS

Motivated by the strong correlation between loudness and human subjective evaluation of English speech using segmental differences between English native speakers and learners, we proposed an objective way to evaluate English learners' timing control by using a measure reflecting perceptual characteristics. This measure was obtained by the time-loudness marker model proposed for the evaluation of synthesized speech. Experiments on the objective evaluation of learners' timing control showed a stronger correlation coefficient of 0.72 than that of 0.54 obtained by the segmental duration differences in a longer sentence set without perceptual weighting. This strong correlation suggests the usefulness of auditory perceptual characteristics in improving the efficiency of objective evaluation. We intend to find more effective measures for objective evaluation of timing control.

6. ACKNOWLEDGEMENTS

This research was conducted in part under the Waseda University RISE research project called "Analysis and modeling of human mechanism in speech and language processing," and it was supported in part by the Grant-in-Aid for Scientific Research (B) No. 20300069, JSPS.

7. REFERENCES

[1] Minematsu, N. "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances." Proc. ICSLP, 1669-1672. 2004.

[2] Tsubota, Y., Dantsuji, M., and Kawahara, T. "Practical use of English pronunciation system for Japanese students in the CALL classroom." Proc. ICSLP, 1689-1692. 2004.

[3] Nakagawa, S., Nakamura, N., and Mori, K. "A statistical method of evaluating pronunciation proficiency for English words spoken by Japanese." Proc. EUROSPEECH, 3193-3196. 2003.

[4] Raux A., and Kawahara. T. "Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning." Proc. ICSLP, 737-740. 2002.

[5] Ito, A., Konno, T., Suzuki, M., and Makino, S. "Improvement of Automatic English Prosody Evaluation Based on Word Clustering Using a Decision Tree." Trans. IEICE, Vol. J91-D, No. 2, 358-366. 2008.

[6] Ito, A., Nagasawa, T., Ogasawara, H., Suzuki, M., and Makino, S. "Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm." Educational Technology Research, vol. 29, 13-23. 2006.

[7] 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. ICPhS, 1733-1736. 2007.

[8] Nakamura, S., Tsubaki, H., Kondo, Y., Nakano, M., and Sagisaka, Y. "On the English timing characteristics by learners in various speech units." Proc. Spring Meet. Acoust. Soc. Jpn., 423-424. 2008. (in Japanese.)

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

[10] Kato, H., Tsuzaki, M., and Sagisaka, Y. "A modeling of the objective evaluation for durational rules based on auditory perceptual characteristics." J. Acoust. Soc. Jpn, Vol. 55-11, 752-760. 1999. (in Japanese with English abstract and figure captions.)

[11] Kato, H., Tsuzakii, M., and Sagisaka, Y. "Effects of phoneme class and duration on the acceptability of temporal modifications in speech." J. Acoust. Soc. Am., Vol. 111, 387–400. 2002.

[12] Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., and Makino, S. "Development of English speech database read by Japanese to support CALL research." Proc. ICA, 557-560. 2004.