TEMPO-NORMALIZED MEASUREMENT AND TEST SET DEPENDENCY IN OBJECTIVE EVALUATION OF ENGLISH LEARNERS' TIMING CHARACTERISTICS

Shizuka Nakamura[1,3], Hajime Tsubaki[1,3], Yusuke Kondo[2,3], Michiko Nakano[2,3] and Yoshinori Sagisaka[1,3]

Graduate School of [1]Global Information and Telecommunication Studies, [2]Education, [3]Language and Speech Science Research Laboratories, Waseda University, Japan
shizuka@akane.waseda.jp, hjm-tsubaki@asagi.waseda.jp, yusukekondo@aoni.waseda.jp, nakano@waseda.jp, sagisaka@waseda.jp

ABSTRACT

In this paper, we present experimental results on tempo-normalized measurements and sentence sets for the objective evaluation of English speech timing by Japanese learners. Phone-independent versus phone-dependent tempo normalizations were compared using raw duration differences between English native speakers and Japanese learners. Sentence length was adopted as a criterion to observe the effect of test sentence differences. Through experiments, high correlations between subjective evaluations and duration differences with normalization showed the remarkable advantage of phone-dependent normalization. Large correlation differences between long and short sentences indicated the need for carefully choosing test materials. Subjective evaluation score estimation by linear regression showed better performance using long sentences and duration differences with phone-dependent normalization than a conventional one using all test sentences and duration differences without normalization.

Keywords: timing control, segmental duration, prosody evaluation, second language learning.

1. INTRODUCTION

In second language learning, research efforts have concentrated on the objective evaluation of learners' second language speech proficiency [1, 2]. In these studies, quite high correlations have been reported between learner speech proficiency and objective evaluation scores consisting of many observable factors, such as number of words uttered per minute and average length of pauses. For better objective scoring of native judgment, speech technologies have started to be employed. Many studies have shown the usefulness of speech recognition technology, indicating additional potential contributions to increase scientific understanding [3-5]. However, since most current works focus on segmental quality evaluation, little research effort has evaluated prosodic quality.

Since there is large language-dependency in such timing control factors as the difference between stress-timing and syllable-timing, timing control is one crucial issue in second language learning. Computational modeling and perceptual studies on segmental duration have revealed the existence of multiple constraints in hierarchically different levels [6, 7]. These works suggest the difficulty of directly analyzing second language learner timing characteristics using only learner speech. We have started to analyze learner timing characteristics by measuring duration differences between English native speakers and Japanese learners using identical English sentences [8, 9].

Correlations between duration differences and subjective evaluation scores clearly show the existence of both language-dependent factors and learner reading proficiency. However, correlation coefficients are still low and further detailed analyses are needed not only to obtain high correlation but also to scientifically understand learner timing characteristics. In this paper, we report the analysis results of the effect of sentence length and duration difference measurements using tempo normalization.

2. TEMPO NORMALIZATION FOR MEASUREMENTS OF DURATION DIFFERENCES

In general, the speech duration of second language learners tends to be longer than the corresponding duration by native speakers. In our previous studies [8, 9], this tendency was clearly observed. Since

duration differences caused by tempo differences extended over every speech unit, they may conceal other factors when analyzing only raw duration differences. To cope with this problem, we adopted tempo normalization to analyze duration differences. For tempo normalization, we tested the following two types of normalizations for each test sentence by considering the duration characteristics of each segment.

(a) Phone-independent normalization (PIN)

Sentence duration is normalized without considering phone-specific lengthening and shortening characteristics. That is, normalization factor λ is calculated by the following equation for every sentence:

Japanese learner's sentence length
$= (1+\lambda)*$ English native speaker's sentence length.

(b) Phone-dependent normalization (PDN)

Sentence duration is normalized by considering phone-specific lengthening and shortening characteristics. That is, normalization factor λ is calculated as follows:

Japanese learner's sentence length
$=$ English native speaker's sentence length $+ \lambda \Sigma \sigma i$,

where σi stands for the standard deviation of the i-th constituent phone and the sum is taken over each sentence. After normalization, each phone duration normdur (i) is calculated as follows:

$$\text{normdur}(i) = \text{dur}(i) + \lambda \sigma i.$$

3. SENTENCE SETS FOR OBJECTIVE EVALUATION

In second language learning, it has been empirically observed that learners have greater difficulties with more complex sentences. Ideally, such differences of reading proficiency would be analyzed using various sentences considering possible factors, though this approach requires data collection and data designing efforts. For this first analysis of test set dependencies, we confirmed the existence of sentence set differences using available databases.

Though we do not know what kinds of complexities may affect reading proficiency, we measured the differences from sentence length. For this purpose, we selected sentence sets from the speech database of second language learner's ERJ (an English speech database read by Japanese students [10]). We adopted 436 speech samples from this database. Test sentences made in consideration of learning prosody consisted of four sentence lengths, as shown in Table 1. Japanese students were given prosodic symbols indicating intonation, phrase boundaries, and pause locations in the given texts and requested to practice pronunciation of the given texts before their recordings. In the recordings, learners were asked to read repeatedly until they thought they could speak with correct pronunciation. During practice and recordings, speech samples uttered by English native speakers were not presented as references.

Table 1: Example sentences used for analysis

Sentence length	Text
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.

Japanese speech samples were mainly uttered by university students (58 males and 63 females) with a wide range of reading proficiencies. As English native reference speech samples, we selected 215 speech samples uttered by English language teachers (8 males and 13 females) who speak General American.

4. EXPERIMENTAL SETUP

4.1. Phone labeling

Phone labeling was performed automatically using the HTK tool. After phone alignment, phone segments were manually examined by experienced researchers with significant knowledge of English phonetics.

4.2. Subjective evaluation by English language teachers

For the analysis, we asked five raters to evaluate all speech samples. All raters have knowledge of English phonetics and careers in teaching English to Japanese learners. Subjective evaluation scores are on a 7-point scale of naturalness in English timing control. It allowed raters to listen to each English speech sample by Japanese learners multiple times.

Table 2: Correlations of subjective evaluation scores with each type of duration difference measurement

Speech unit	With normalization		Without normalization
	PIN	PDN	
Phoneme			
Phoneme	-0.35	-0.44	-0.37
Vowel	-0.38	-0.41	-0.41
Strong vowel	-0.31	-0.20	-0.25
Weak vowel	-0.32	-0.43	-0.40
Vowel in content word	-0.38	-0.31	-0.31
Vowel in function word	-0.17	-0.40	-0.33
Consonant	-0.20	-0.29	-0.20
Voiced consonant	-0.18	-0.29	-0.18
Unvoiced consonant	-0.12	-0.20	-0.13
Consonant in content word	-0.20	-0.29	-0.18
Consonant in function word	-0.15	-0.23	-0.18
Syllable			
Syllable	-0.33	-0.39	-0.27
Stressed syllable	-0.31	-0.29	-0.19
Unstressed syllable	-0.23	-0.38	-0.32
Open syllable	-0.27	-0.44	-0.33
Closed syllable	-0.30	-0.33	-0.23
Word			
Word	-0.33	-0.39	-0.31
Content word	-0.34	-0.34	-0.19
Function word	-0.20	-0.35	-0.33
Pause			
Pause	-0.33	-0.35	-0.36

5. CORRELATION ANALYSIS

5.1. Correlations between duration differences and subjective evaluation scores

After tempo normalization, we measured the duration differences between English native speakers and Japanese learners not only in every speech unit that should have linguistically reasonable correlations with subjective evaluation scores but also its referential speech unit. We got correlation coefficients between duration differences and subjective evaluation scores. The results are shown in Table 2. For comparison, raw duration differences without normalization were also calculated.

As shown in Table 2, duration differences with tempo normalization show higher correlations than those without normalization in almost all speech units. Though opposite results are observed in some speech units such as strong vowels, these are caused by the contribution of speech rates. In particular, phone-dependent normalization (PDN) shows higher correlations than phone-independent normalization (PIN). These results show PDN's usefulness in duration difference measurements to explain subjective evaluation characteristics.

Table 3: Increasing Japanese learner sentence duration ratios to those of English native speakers

Sentence length	Sentence duration ratio
Very short (VS)	1.18
Short (S)	1.22
Long (L)	1.23
Very long (VL)	1.39

Table 4: Correlation dependence on sentence length

Speech unit	With PDN		Without normalization	
	VL	VS	VL	VS
Phoneme				
Phoneme	-0.50	-0.21	-0.37	-0.36
Vowel	-0.52	-0.27	-0.44	-0.40
Strong vowel	-0.19	-0.15	-0.15	-0.28
Weak vowel	-0.61	-0.17	-0.42	-0.26
Vowel in content word	-0.33	-0.23	-0.24	-0.36
Vowel in function word	-0.57	-0.13	-0.40	-0.12
Consonant	-0.36	-0.04	-0.13	-0.06
Voiced consonant	-0.34	-0.10	-0.03	-0.11
Unvoiced consonant	-0.31	0.00	-0.14	0.00
Consonant in content word	-0.37	-0.05	-0.08	-0.08
Consonant in function word	-0.27	-0.05	-0.15	-0.04
Syllable				
Syllable	-0.48	-0.21	-0.32	-0.07
Stressed syllable	-0.33	-0.10	-0.12	0.01
Unstressed syllable	-0.47	-0.18	-0.45	-0.22
Open syllable	-0.54	-0.35	-0.37	-0.10
Closed syllable	-0.40	-0.09	-0.26	-0.08
Word				
Word	-0.48	-0.21	-0.37	-0.06
Content word	-0.40	-0.17	-0.22	0.01
Function word	-0.42	-0.14	-0.44	-0.12
Pause				
Pause	-0.49	-0.11	-0.41	-0.11

5.2. Timing dependence on sentence length

First, we measured the sentence duration lengthening characteristics in relation to sentence length. Table 3 shows the Japanese learners' average sentence duration ratios to those of English native speakers. In every sentence length, ratios are

Table 5: Increasing correlations caused by effects of sentence lengths and tempo normalization

Test condition	Sentence length	Duration measure	Closed	Open
PDNL	VL	With PDN	0.61	0.57
NONL		Without normalization	0.61	-0.06
PDNS	VS	With PDN	0.39	0.01
NONS		Without normalization	0.26	0.04

bigger than 1.0, which means that duration is lengthened. The lengthening ratio becomes larger when sentence length becomes longer.

Next, we calculated the correlations between subjective evaluation scores and duration differences between English native speakers and Japanese learners. As shown in Table 4, higher correlations were obtained for very long sentence sets than very short ones in both duration differences with and without normalization. For this calculation, we adopted PDN since it showed higher correlations than PIN in the previous section.

6. PREDICTION OF PROFICIENCIES IN ENGLISH TIMING CONTROL

To predict subjective evaluations of timing control naturalness, a linear regression model was adopted. To compare the prediction performance of different measurements and test sets, models were trained in four different conditions. Two were very long or very short sentence sets with phone-dependent normalization (PDNL, PDNS). The other two were very long or very short sentence sets without normalization (NONL, NONS).

111 samples were used in each condition. For model training, four-fifths of the samples were used, and the rest of the samples were used for the test. Table 5 shows the correlation coefficients between subjective evaluation scores and predicted scores. As shown in Table 5, a higher correlation coefficient of 0.57 was obtained for the test set in PDNL, which included very long sentence sets using duration differences with phone-dependent normalization. This value is remarkably higher than correlation coefficients in other conditions.

7. CONCLUSION

We analyzed tempo-normalized measurements and test set dependency in the prediction of subjective evaluation scores of English timing characteristics by objective measurements. Correlation analyses and subjective evaluation experiments showed the effectiveness of tempo normalization and longer sentence sets. These facts suggest the possibility of better prediction by looking for much finer measurements and measurement targets. We would like to continue to search for new measurements and differences in test targets to obtain a better prediction model and scientific understanding of the underlying control mechanism.

8. ACKNOWLEDGEMENTS

This research was conducted in part under the Waseda University RISE research project of "Analysis and modeling of human mechanism in speech and language processing" and supported in part by the Grant-in-Aid for Scientific Research A, No. 16200016 of JSPS.

9. REFERENCES

[1] Nakano, M., Kondo, Y., Ueda, N., Tsutsui, E., Owada, K. 2006. A study of evaluation in the speaking ability of Asian learners of English by using FACETS. Proc. the 4th Annual Conference of the Japan Association for Research on Testing Japan, 38-41.

[2] Bernstein, J., Barbier, I., Rosenfeld, E., Jong, John H.A.L. de. 2004. Theory and data in spoken language assessment. Proc. INTERSPEECH 2004 Jeju, 1685-1688.

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

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

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

[6] Sagisaka, Y. 2003. Modeling and perception of temporal characteristics in speech. Proc. ICPhS 2003 Barcelona, 1-6.

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

[8] Muto, M., Sagisaka, Y., Naito, T., Maeki, D., Kondo, A., Shirai, K. 2003. Corpus-based modeling of naturalness estimation in timing control for non-native speech. Proc. EUROSPEECH 2003 Geneva, 498-501.

[9] Nakamura, S., Tsubaki, H., Sagisaka, Y. 2007. On the measurement of English timing characteristics of Japanese learners. Proc. Spring Meet. Acoust. Soc. Jpn. Tokyo, 249-250.

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