Acoust. Sci. & Tech. **31**, 4 (2010)

PAPER

Model-based automatic evaluation of second-language learner's English segmental duration characteristics

Chatchawarn Hansakunbuntheung[1,*], Hiroaki Kato[2] and Yoshinori Sagisaka[1,†]

[1]*GITI/Language and Speech Science Research Laboratories, Waseda University, 29–7 Building, 1–3–10 Nishi-Waseda, Shinjuku-ku, Tokyo, 169–0051 Japan*
[2]*National Institute of Information and Communications Technology/ ATR Media Information Science Laboratories, 2–2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0288 Japan*

(*Received 23 July 2009, Accepted for publication 13 January 2010*)

Abstract: In this paper, we propose a method of automatically measuring the segmental duration characteristics of a second-language learner's speech as a means to evaluate language proficiency in terms of speech production. We propose the use of duration differences from native speakers' speech as an objective evaluation score to evaluate the learner's English segmental duration characteristics. To provide flexible evaluation without the need to collect any additional native-English reference speech, we employed predicted normalized segmental durations using a statistical duration model instead of measured raw durations of native reference speech. The proposed evaluation method was tested using English speech data uttered by multiple Thai-native learners' groups with different amounts of experience of English study in English-as-an-official-language countries. An evaluation experiment showed that the proposed measure based on duration differences is strongly correlated with the amount of English study. Moreover, segmental duration differences revealed Thai learners' speech-control characteristics such as stress assignment on word-final syllables. These results support the effectiveness of the proposed model-based objective evaluation.

Keywords: Segmental duration, Automatic evaluation, Second-language learning, Linear regression model

PACS number: 43.72.Ar, 43.70Kv, 43.71.Gv [doi:10.1250/ast.31.267]

1. INTRODUCTION

Spoken-language learning is a bidirectional process that requires evaluative feedback to identify and describe a learner's disfluencies and spoken errors with the aim of improving the learner's speaking proficiency. As feedback, learners also require language proficiency measures to characterize their own current language proficiency levels, which allow them to monitor their progress. However, proficiency scores alone do not provide sufficient feedback for language learners to pinpoint their speaking flaws. We need more informative feedback that can identify a learner's weak points in speaking. Furthermore, if this information can be automatically obtained, learners can evaluate themselves and keep track of their proficiency anytime without the need for a human evaluator.

The existing conventional language-proficiency evaluations generally provide subjective feedback by professional human evaluators. To standardize the evaluation among evaluators, several language-proficiency evaluation frameworks for speaking have been established, such as the Common European Framework of Reference (CEFR) for Languages: Learning, Teaching, Assessment [1], Interagency Language Roundtable (ILR) [2], American Council for the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines for Speaking and the Oral Proficiency Interview (OPI) [3], and, the ACTFL-ALC-based Standard Speaking Test (SST) of English for Japanese speakers [4]. These frameworks use various aspects of spoken features to characterize specific components of communicative competence [5–7], such as the mastery of pronunciation, fluency, prosodic, lexical, grammatical, and pragmatic subskills, and to evaluate learners' speaking proficiency, such as spoken-speech characteristics (pronunciation, rhythm, intonation, speech rate, pause structure, fluidity),

*e-mail: chatchawarn.hansakunbuntheung@nectec.or.th
†e-mail: ysagisaka@gmail.com

language use (vocabulary and grammar), and topic development (content and coherence).

However, various problems of subjective evaluation have arisen and remain unresolved, such as the time-consuming nature of manual evaluation, inconsistency among evaluators, evaluators' different and subjective perspectives, and the need for multiple evaluators to reduce evaluator subjectivity [8]. Furthermore, the increasing number of learners has created a greater need for professional evaluators as well as more time devoted to evaluation. Therefore, automatic evaluation methods based on objective measures have been proposed to solve these problems.

With the progress of automatic speech recognition (ASR), many research studies have been conducted to achieve automatic evaluations of learner proficiency using ASR-based pronunciation assessment [9,10]. To investigate the reliability of automatic evaluations, the correlations between human and various machine scores (i.e., the hidden Markov model (HMM) log-likelihood, posterior-probability, phone recognition accuracy, segment duration, and syllabic timing scores) were compared [11] and it was found that the posterior-probability-based and speech-rate-normalized-duration scores have a higher correlation with human ratings than the other scores. In addition to the pronunciation scores, the speech rate and fluency features were studied and the following results were obtained: 1) the speech rate was the best measure among the examined measures, 2) native speakers were more fluent than nonnatives, and 3) the temporal measures were significantly different between the two groups [12–15]. More recently, a number of studies focused on the use of temporal, prosodic, and fluency measures and partly succeeded in evaluating a second-language (L2) learner's speech [16,17]. As a more sophisticated and interpretable method, an automatic speaking test called SET-10 used linguistic content, pronunciation, and phonological fluency as measures for predicting intelligibility, and the results demonstrated the comparability of human ratings with the automatic measures [18]. A research team at the Educational Testing Service (ETS) adopted a classification tree using the pace and clarity of speech to evaluate the features of training systems for the Test of English as a Foreign Language (TOEFL) [19–22]. These studies can partly describe which factors human evaluators use to rate learner proficiency.

By using these methods of automatic evaluation, interactive tests can be developed to provide immediate scoring feedback to language learners. Nevertheless, these evaluations still do not clearly describe precise quantitative factors that a learner improperly produces in speaking. These factors and their feedback are necessary information for learners to correct their speaking skills. Consequently, this raises the issue of quantitative language-proficiency evaluation for speaking, which is crucial for improving the tools and systems used in language learning. Furthermore, this issue will particularly benefit self-learning language learners.

In the proficiency-evaluation frameworks [1–3,23], pronunciation is considered the basic criterion for discriminating between beginner and higher levels. For higher levels, prosodic characteristics are used as criteria to evaluate a wider range of proficiency, especially to distinguish between advanced and lower levels within this range. As emphasized in many frameworks for language-proficiency evaluation, duration is an essential issue. Furthermore, duration is a fundamental acoustic property of speech that can be directly measured from speech. Thus, in this paper we focus on duration-based language learning evaluation.

In this work, we propose an English proficiency-evaluation scheme based on an English duration model along with a duration-difference-based measure and consequently provide an objective evaluation score. More specifically, the evaluation scheme measures the duration differences between an individual learner's segmental durations and those of native speakers that were produced from the English duration model, and uses these differences to provide an evaluation score. The use of a duration model enables a flexible choice of test sentences without requiring any additional or identical speech corpus of native samples for comparison. To test the model's effectiveness, we have applied it to English speech uttered by multiple groups of Thai learners having different amounts of experience of English study. In Section 2, we first introduce proficiency evaluation using segmental duration differences and a statistical segmental duration model of native English, and discuss the objective measurement of duration differences. Next, in Section 3, we explain our experimental setup consisting of speech corpora of multiple groups of speakers with different English study experiences. In Section 4, experimental details and results are presented. Finally, in Section 5, the conclusions of this paper are given.

2. MODEL-BASED PROFICIENCY EVALUATION AND OBJECTIVE MEASURES

2.1. Model-Based Proficiency Evaluation

Figure 1 shows an overview of the proposed evaluation scheme. The main concept for evaluating a learner's English proficiency based on segmental duration control differences is that we calculate an objective measure representing the average difference between the segmental durations of an individual target learner and those of native speakers as an alternative to conventional subjective measures. To be free from the requirement of an exhaustive

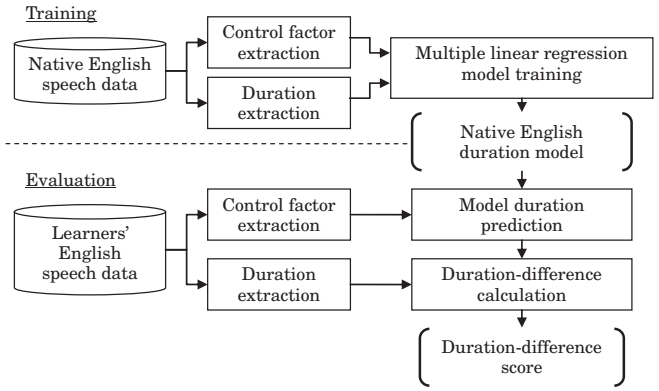

Fig. 1 Overview of the model-based automatic evaluation of second-language learners' English segmental duration characteristics.

collection of native speakers as references, a duration model of native English was statistically built using native English speech data uttered by multiple native speakers. The original segmental durations and a set of duration control factors were then extracted from the native speech data for modeling. Consequently, the use of a duration model statistically represents the average durational characteristics of native English speakers by normalizing the characteristics of the individuals. Then, we used this duration model to predict a representative set of native segmental durations for assessing content. Next, we measured the duration characteristics and duration differences between the observed durations and those statistically predicted from the model to compare the differences between native speakers and learners in English from the viewpoint of segmental duration control.

To compute the statistical segmental durations of the average durational characteristics of native English speakers, we adopted segmental durations normalized by the speech rate. Before modeling, we normalized the segmental duration of each phone for each speaker using z-score normalization with the mean and standard deviation (SD) to eliminate any effect of speech rate for interspeaker comparison. The mean and standard deviation used here are speaker-dependent phone-independent values calculated from all of the phones in the speech data. In this paper, we used the inverse of this mean as the speaker-dependent speech rate for analysis. For the modeling, we adopted multiple linear regression based on categorical factors [24] as shown in Eq. (1).

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc}\delta_{fc}(i) \quad :i = 1, 2, 3, \ldots, N \tag{1}$$

$$\delta_{fc}(i) = \begin{cases} 1 & :\text{if the } i\text{th speech segment falls into} \\ & \text{category } c \text{ of factor } f, \\ 0 & :\text{otherwise} \end{cases} \tag{2}$$

Here N, $\hat{y}_i$, $\bar{y}$, x_{fc}, and $\delta_{fc}(i)$ represent the number of items of data, the predicted duration of the ith speech segment, the mean duration of all samples, the regression coefficient of category c of control factor f, and the characteristic function, respectively. To feed data into the model, each category c of factor f of the ith speech segment was encoded using the characteristic function $\delta_{fc}(i)$. By adopting the least-squares error minimization technique, modeling coefficients representing the contributions of the control factors were calculated.

As shown in Table 1, for the control factors, we employed the current and four context phones [25]; phone positions in each syllable, word, and phrase; the numbers of constituent phones in each syllable, word, and phrase; syllabic stress; syllable positions in each word and phrase; the numbers of constituent syllables in each word and phrase; and the general and specific parts of speech [26]. These factors were adopted by referring to our previous study on English segmental duration [27]. However, we do not claim that these factors provide a full characterization of communicative competence.

2.2. Objective Duration-Difference Measures for Proficiency Evaluation

To evaluate learner speaking proficiency based on an English duration model, we compared the deviations of actual speech durations from the predicted durations of two speaker groups, i.e., English native speakers and Thai learners.

We measured phone durations from segmented speech data. Then, we used the English duration model with the control factors from the speech data to predict native English segmental durations. Next, the average difference between the measured and estimated reference durations was calculated for each speaker using the root-mean-square (RMS) value. By considering duration differences as an evaluation measure, we expect to discover some durational parameters representing different elasticity characteristics among the speakers' native languages. Furthermore, we calculated not only the deviations from predicted durations but also the average speech rate and the standard deviation of observed phone durations for the two speaker groups. The speech rate and the standard deviation of phone durations are equal to the values used for the duration normalization explained in Subsection 2.1.

3. EXPERIMENTAL SPEECH DATA

We employed three types of English speech databases. The first one was the Carnegie Mellon University (CMU) ARCTIC database [28], which consisted of sentences read by native English speakers. This database was used for segmental duration modeling and to evaluate the model's

Table 1 Control factors and categories employed in linear regression modeling of normalized segmental duration of native English. (Note: For the "phone position in word," "phone position in phrase," "syllable position in word," and "syllable position in phrase" factors, the labels "I," "AI," "M," "BF," and "F" represent the initial, after-initial, mid-, before-final, and final syllable positions in a polysyllabic word or phrase, respectively. The label "S" represents the phone or syllable in a monosyllabic word or phrase. The indices of the positions represent the word or phrase length in syllables. If not defined, they represent positions in polysyllabic words with 4 syllables or more, and the labels "AI" and "BF" denote the second and second-to-last syllable positions, respectively.)

Factor	Category
Current phone	CMU's English phones [25]: aa, ae, ah, ao, aw, ay, b, ch, d, dh, eh, er, ey, f, g, hh, ih, iy, jh, k, l, m, n, ng, ow, oy, p, r, s, sh, t, th, uh, uw, v, w, y, z, zh
Phone before phone preceding current phone	CMU's English phones [25] and pause
Phone preceding current phone	
Phone following current phone	
Phone following phone after current phone	
Phone position in syllable (current phone position m, number of constituent phones in syllable n)	(m, n): $m = 1, \ldots, n$ and $n = 2, \ldots, 7$
Numbers of constituent phones in syllable	$1, \ldots, 7$
Phone position in word	I2, F2, I3, M3, F3, I, AI, M, BF, F, or S
Numbers of constituent phones in word	1, 2, 3, 4, 5, 6, 7, 8, 9–10, 11–12, 13–14
Phone position in phrase	I2, F2, I3, M3, F3, I, AI, M, BF, F, or S
Numbers of constituent phones in phrase (grouped by three)	1–3, 4–6, 7–9, . . . , 73–75, 76–78
Syllabic lexical stress	Stressed, Unstressed
Syllable position in word	I2, F2, I3, M3, F3, I, AI, M, BF, F, or S
Numbers of constituent syllables in word	$1, \ldots, 7$
Syllable position in phrase	I2, F2, I3, M3, F3, I, AI, M, BF, F, or S
Numbers of constituent syllables in phrase	$1, \ldots, 30$
General part of speech	Function word, Content word
Specific part of speech	cc, cd, dt, ex, fw, in, jj, jjr, jjs, md, nn, nnp, nnps, nns, of, pdt, pos, prp, rb, rbr, rbs, rp, sym, to, uh, vb, vbd, vbg, vbn, vbp, vbz, wdt, wp, wrb [26]

accuracy in reflecting native duration characteristics. The database was separated into two phonetically balanced sets: set A with 593 sentences and set B with 539 sentences. The contents of these two sets were completely different. The second database was a read English speech database consisting of the fairy tale "The North Wind and the Sun," from the Chinese University of Hong Kong (CUHK) Chinese Learners of English Speech Corpus (CUCHLOE) [29,30], which was used to test the effectiveness of the proposed evaluation scheme. The sentences were uttered by native English speakers and speakers from the English-as-an-official-language countries of USA, UK, Canada, Hong Kong, and India. The third database was also a test-speech database collected at the National Electronics and Computer Technology Center, Thailand (NECTEC). This database contains speech data consisting of sentences from the same fairy tale uttered by 45 Thai learners with different English study backgrounds and one Indian English speaker.

These three speech databases were divided into four groups, which were used for modeling and analysis. The first group consisted of CMU ARCTIC set A uttered by four US speakers [28]. This was used as the training set for the prediction of segmental durations by reference native speakers. The second group consisted of CMU ARCTIC set B, uttered by the same four speakers. We refer to this group as a closed-speaker open-text set, which was used to evaluate the consistency of the model. If our evaluation scheme can be effectively used to calculate the duration differences between learners by a model-based approach, the predicted duration difference of the second group from the first group (training set) is expected to be smaller than that of the other groups, which use different speakers from those in the training (open-speaker sets).

Table 2 Thai learners categorized by their amount of experience of English study.

Period of English study in English-as-an-official-language countries (Years)	Number of learners (Persons)
<1	34
1–5	3
6–10	3
>10	5

Table 3 Rating agreement among evaluators' scores.

Evaluator	Average correlation coefficient
Native	0.82
Nonnative English teacher #1	0.80
Nonnative English teacher #2	0.82
Nonnative English teacher #3	0.85
Nonnative English teacher #4	0.85
Nonnative English teacher #5	0.85
Nonnative English teacher #6	0.79

Table 4 Correlation coefficients between the subjective scores and the proposed duration-difference scores for learners' data using different segmentations (sample size: 45 Thai learners).

	Correlation coefficient	p-value
Duration differences (proposed) using automatically segmented data	−0.62	0.0000
Duration differences (proposed) using manually segmented data	−0.58	0.0000

To evaluate the model's validity with various English accents, we used an open-speaker open-text set as the third group. It included the data obtained from three non-US-accent English speakers from CMU ARCTIC set B, six speakers from CUCHLOE, and one speaker from NECTEC. The fourth group contained the speech data of 45 Thai learners of English from NECTEC with various amounts of experience of English study. We used this last group as a test set to evaluate the English segmental duration characteristics of Thai learners. The learners were further categorized into four subgroups based on the amount of time spent in an English-as-an-official-language country as shown in Table 2.

Then, the speech data for evaluation were segmented by an HMM-based automatic segmentation scheme. We adopted HMM Toolkit (HTK) using acoustic models based on the VoxForge English speech corpus [31]. To reduce segmentation errors due to the incorrect pronunciation of nonnative speakers, we used a conventional HMM-based forced-alignment segmentation scheme with speaker adaptation based on acoustic models and pronunciation variation at the phone level. The acoustic models were adapted for each speaker by using the evaluation speech data obtained from their own voices.

4. EXPERIMENTAL RESULTS

4.1. Evaluation of Performance of Proposed Method

To evaluate the performance of the proposed method, we carried out a subjective evaluation using human evaluators. Briefly, we asked a group of volunteers to listen to the speech data we used in this paper. Then, the volunteers assigned a "similarity to native speech" score (an overall score considering naturalness, fluency, accent, stressing, articulation, etc.) to each speaker. The score has a 9-degree scale (5 major degrees and 4 minor degrees) i.e., 1-Very poor (unintelligible), 1.5, 2-Poor, 2.5, 3-Medium, 3.5, 4-Good, 4.5, 5-Very good (nativelike).

Seven evaluators took part in this test. One of them was a native speaker of American English. The other six were not native English speakers but native Japanese speakers who had substantial experience in the field of English education such as English teaching. We first investigated the agreement among the evaluators to test the validity of using nonnative evaluators. Table 3 shows the average correlation coefficients between the scores of each evaluator and the other evaluators. They are equal to or more than 0.79, indicating high rating agreement among the evaluators. Furthermore, the high correlation coefficient between the scores of the native evaluator and the other nonnative evaluators of 0.82 indicates that the ratings given by the native and nonnative evaluators are comparable.

Next, we calculated the average correlation coefficients between the subjective scores given by the evaluators and the proposed duration-difference measures based on the automatically segmented speech data. As shown in Table 4, the proposed method had a high absolute correlation coefficient of −0.62 for the open-speaker learner set. This high value implies the similarity between the subjective scores determined by human evaluators and the proposed scores calculated by the proposed method. Thus, these results indicate the satisfactory evaluation performance of the proposed method and the proposed evaluation score.

Then, to test the reliability of the automatic segmentation scheme, we compared the evaluation performance using automatically segmented data with that using manually segmented data. For the test set including open-speaker native speakers and learners, we found a high correlation coefficient ($r = 0.76$, $p < 0.05$) between results obtained using automatically and manually segmented speech data. Furthermore, as shown in Table 4, we also

Table 5 Comparison of evaluation performance between the proposed and conventional methods based on correlation coefficient between scores for learner speech data assigned by human evaluators and automatic evaluation method.

	Evaluation score	Correlation coefficient (maximum absolute)
Proposed	- Phone duration differences between those of native English duration model and measured durations	0.62
Ito *et al.* (2006) [32]	- Intonation based: Normalized and gradient F_0 of intonation phrase	0.53
	- Rhythm based: Relative duration and PIER of phone, word, and phrase	0.36
Yamashita *et al.* (2004) [33]	- F_0 distances, gradients, and approximation error of F_0 fitting based on word boundaries	0.49
	- Power	0.38
	- Utterance duration	0.44
	- All the above	0.51

calculated the correlation coefficient between the subjective scores and the duration-difference scores using manually and automatically segmented data for the open-speaker learner set to compare their evaluation performances. The correlation coefficients for both methods of segmentation in Table 4 indicate similar evaluation performances. Actually, the proposed method using the automatically segmented data exhibited a slight improvement over that using the manually segmented data. Thus, all these results show that the proposed method using automatically segmented data exhibited an evaluation performance comparable to that using manually segmented data.

Moreover, we compared the evaluation performance between the proposed method and other conventional methods [32,33] as shown in Table 5. The results showed that the proposed method gave a higher correlation with the scores of human' than the conventional methods. This superiority should be particularly emphasized if we consider that the proposed method only used one aspect of speech properties, i.e., duration, while the conventional methods used multiple speech properties.

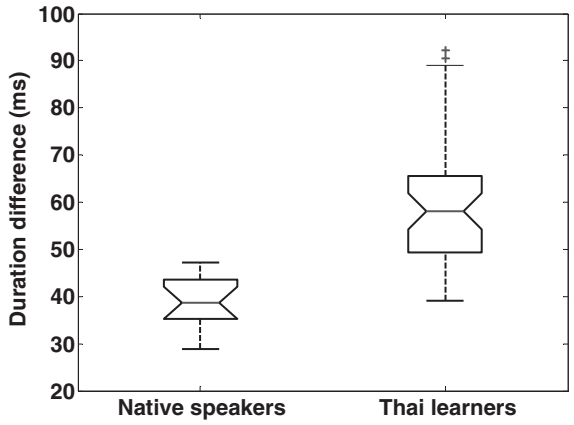

Fig. 2 Comparison of RMS duration differences from predicted durations between native English speakers and Thai learners. In this and subsequent box plots, the horizontal line at the center of each box indicates the median, the vertical length of the box indicates the interquartile range (the range between the lower and upper quartiles), each whisker indicates the data point furthest from the edge of the box excluding those further than 1.5 times the length of the box, and the plus symbols indicate outliers that are further out than the ends of the whiskers.

4.2. Correlation between Objective Duration-Difference Measures and Amount of English Study Experience

Figure 2 shows a comparison of the RMS duration differences between the predicted durations of native English speakers and Thai learners. As the figure shows, the Thai learners produced larger deviations from the English duration model (the median and mean of the prediction difference are 58.1 and 59.2 ms, respectively) than the native English speakers (the median and mean of the prediction difference are 38.6 and 38.7 ms, respectively). Furthermore, the central quartiles of the distributions for the native English and Thai learner groups are clearly separated. This result suggests the usability of the duration differences from the predicted durations for quantifying the differences between native speakers and Thai learners.

Interestingly, similar differences between native English speakers and Thai learners were found by observing the standard deviation of phone durations as shown in Fig. 3. We performed the Pearson product-moment correlation test (significant level, $\alpha = 0.05$) and found a high correlation of 0.98 between the differences from the predicted durations and the standard deviation of phone durations as shown in Fig. 4. As mentioned earlier, a number of previous studies reported the validity of the speech-rate measure [12–15]. However, the analysis in the current study suggested that this was not the case. As shown in Fig. 5, the difference in the speech rate between

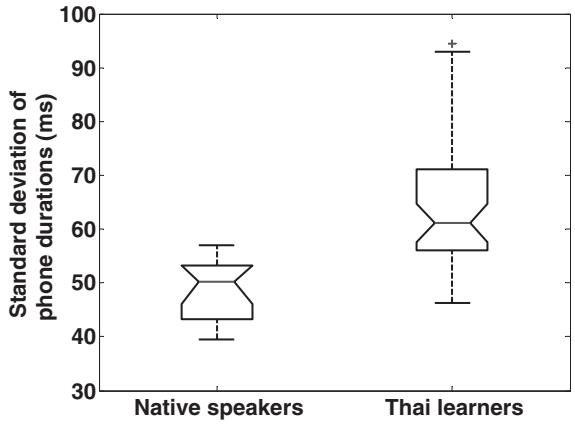

Fig. 3 Comparison of standard deviation of phone durations between native English speakers and Thai learners.

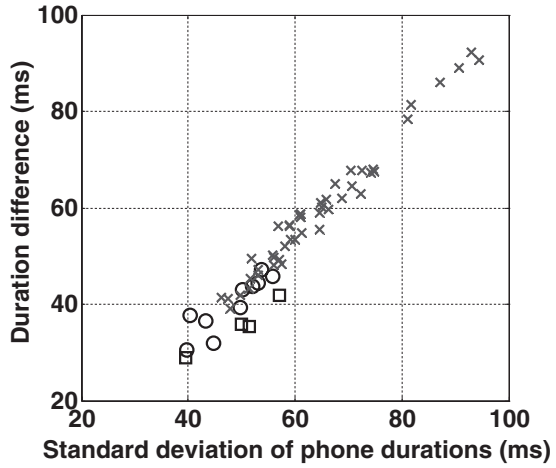

Fig. 4 Comparison between RMS duration differences from predicted durations and standard deviations of phone durations ('□,' '○,' and '×' represent native closed speakers, native open speakers, and Thai learners, respectively).

native speakers and Thai learners was very small. The average speech rates of the native speaker and Thai learner groups were 12.3 and 11.1 phone/s, respectively.

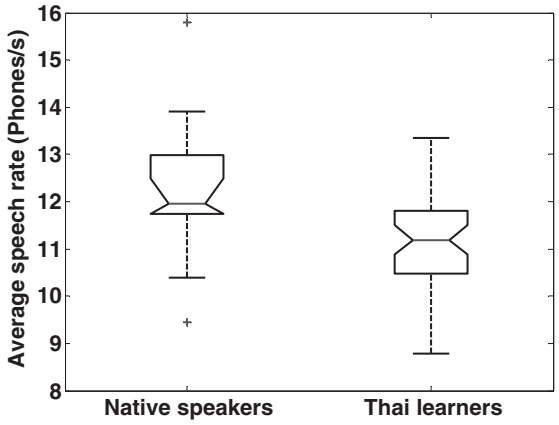

Fig. 5 Comparison of speech rates between native English speakers and Thai learners.

From Figs. 2, 3, and 5, we performed further statistical similarity tests between native speaker and Thai learner groups using the two-sample *t*-test (two-tailed test at a significance level of $\alpha = 0.05$) on the three above-mentioned measures, i.e., duration difference, standard deviation of phone duration, and speech rate. As shown in Table 6, all the observed measures gave statistically significant differences between native speaker and Thai learner groups (significance values $p_{Duration\_Difference} = 1.0 \times 10^{-10}$, $p_{Phone\_Duration\_SD} = 5.4 \times 10^{-8}$, and $p_{Speech\_Rate} = 3.2 \times 10^{-3}$). The obtained significance values suggested that the duration-difference measure was more effective in separating native speaker and Thai learner groups than the SD of phone duration and speech rate. To investigate the classification accuracy of the measures, we performed a linear discrimination analysis (LDA) on native speaker and Thai learner groups with three different measures. The LDA used Fisher discriminant coefficients and assigned equal prior probabilities to both native speaker and learner groups for initialization. Then, we evaluated the accuracy of the observed measures based on the LDA using leave-one-out cross-validation. Table 7 shows the accuracy of the three measures using the

Table 6 Results of two-sample *t*-test to evaluate similarity between native and learner score distributions using different measures.

Measures	Levene's test for equality of variances		*t*-test for equality of means		
	F	*p*	*t*	*df*	*p* (two-tailed)
Duration difference ($\sigma^2_{\text{Native}} \neq \sigma^2_{\text{Learner}}$ assumed)	4.874	0.031	−8.115	50.586	0.000 (1.0×10^{-10})
SD of phone duration ($\sigma^2_{\text{Native}} \neq \sigma^2_{\text{Learner}}$ assumed)	4.747	0.034	−6.503	45.442	0.000 (5.4×10^{-8})
Speech rate ($\sigma^2_{\text{Native}} = \sigma^2_{\text{Learner}}$ assumed)	1.127	0.293	3.084	57	0.003 (3.2×10^{-3})

Table 7 Linear discrimination analysis using Fisher discriminant coefficient for native speaker and learner groups with three different measures.

Measure	Accuracy (%)	Unit	Group	Predicted group membership		
				Native	Learner	Total
Duration difference	83.1	Count	Native	14	0	14
			Learner	10	35	45
		%	Native	100.0	0.0	100.0
			Learner	22.2	77.8	100.0
SD of phone durations	78.0	Count	Native	13	1	14
			Learner	12	33	45
		%	Native	92.9	7.1	100.0
			Learner	26.7	73.3	100.0
Speech rate	71.2	Count	Native	10	4	14
			Learner	13	32	45
		%	Native	71.4	28.6	100.0
			Learner	28.9	71.1	100.0

LDA. These results clearly show that the duration-difference measure outperformed the other measures in discrimination between native speakers and Thai learners.

To examine the relationship between the differences from predicted durations and the amount of English-study experience more closely, we compared the differences from predicted durations between native English speakers and Thai learners grouped by periods of years of English-study experience in English-as-an-official-language countries such as USA, UK, Australia, and India. As shown in Fig. 6, noticeable duration differences were observed between learners groups depending on the time spent studying English overseas. The duration of the closed-speaker open-text set showed the least difference from that of the training set (i.e. the closed-speaker closed-text set). This group also showed the smallest duration differences among all speaker groups. Accordingly, the results showed the consistency and reasonable prediction accuracy of the model both for the training set and for the open set.

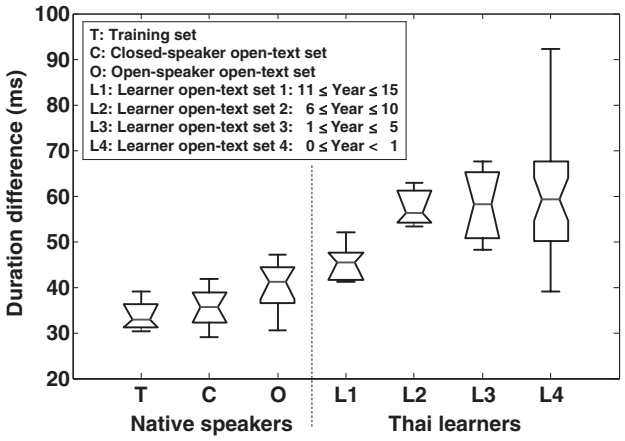

Fig. 6 Comparison of RMS duration differences from predicted durations between native English speakers (C: closed speakers, O: open speakers) and Thai learners (L1–L4), grouped by periods of years of English-study experience in English-as-an-official-language countries.

For the other-accented native speakers in the open-speaker data set, their duration differences were much closer to those of speakers in the training set and smaller than most of those for the Thai learners. Interestingly, the learners living in English-as-an-official-language countries for more than 10 years showed a clear decrease in the distance from the reference model, while the learners with less than 10 years of experience in such countries showed larger duration differences with a larger variation in phone duration differences than more experienced learners. Regarding the correlation between the period of time in English-as-an-official-language countries and the duration difference, the results show a negative correlation coefficient (significance level, $\alpha = 0.05$) of -0.37. These results indicate the effectiveness of the proposed objective measure. Since other learner background factors that have not been taken into account may also affect this measure, wide variations of duration differences in the Thai learner groups, especially in the least experienced group, can be found. In the least experienced group, we found that five speakers from this group had very short-term study experience in English-as-an-official-language countries. Thus, we expect that some effects of this short-term experience caused the overlap of the duration differences

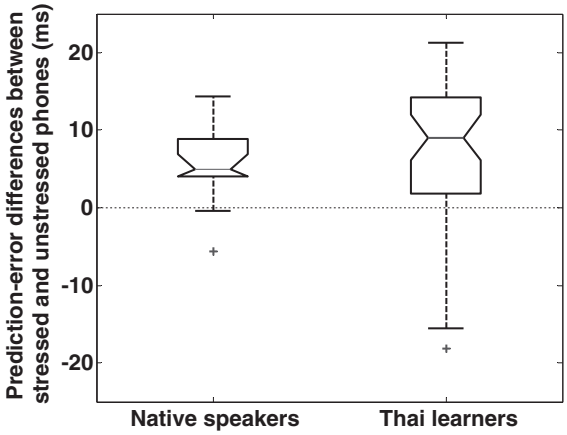

Fig. 7 Comparison between native speakers and learners for prediction-error differences of phone durations between stressed and unstressed syllables. Each plot represents the average of individual prediction-errors in stressed syllables subtracted by that in unstressed syllables.

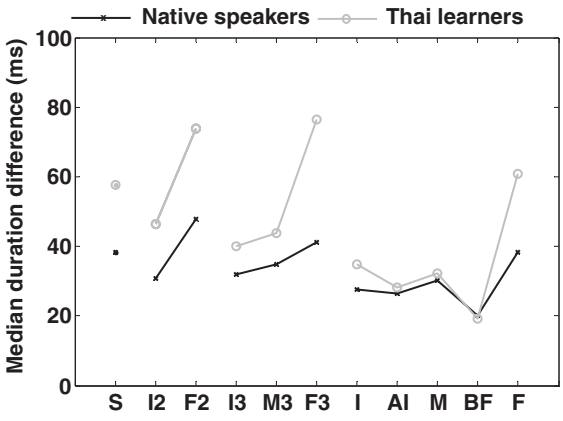

Fig. 8 Median duration differences from predicted durations at different positions in words of different lengths between native English speakers and Thai learners (where the labels "I," "AI," "M," "BF," and "F" represent the initial, after-initial, mid-, before-final, and final syllable positions in a polysyllabic word or phrase, respectively. The label "S" represents the phone or syllable in a monosyllabic word or phrase. The indices of the positions represent the word or phrase length in syllables. If not defined, they represent positions in polysyllabic words with 4 syllables or more, and the labels "AI" and "BF" denote the second and second-to-last syllable positions, respectively.)

between the least experienced and more experienced groups. Thus, these wide variations of duration differences need further investigation in more detail in future works.

4.3. Duration-Difference Analysis to Characterize English Segmental Duration Characteristics of Thai Learners

To further investigate the effectiveness of the proposed duration differences, we analyzed the duration differences between Thai learners and native English speakers. We observed the differences in the deviation of duration caused by each of the model's control factors by contrasting the data of native speakers and Thai learners. Since Thai is a stress-timed language similarly to English, we first analyzed the duration differences caused by English stress. Figure 7 indicates the average of individual duration differences (the proposed prediction-error-based measure) observed in phones in stressed syllables subtracted by that observed in phones in unstressed syllables. As shown in Fig. 7, negative values for individual differences were mostly found in the learners' data. In the case of native English speakers, we found this type of error in two of the non-US speakers. Most of the Thai learners who produced this type of error had no or minimal experience of study in English-as-an-official-language countries. These results suggest that a learner with this negative value tends to use a different control strategy from native speakers to cope with stressing, and a probable cause of the observed negative value is the misplacement of stress on an unstressed syllable.

Furthermore, by analyzing the durational differences at different syllable positions, it was found that the Thai learners always produced larger duration differences at the end of a word or phrase than the native English speakers. Figure 8 shows this characteristic clearly. In the stress placement system of Thai, the primary stress is always located at the last syllable of a word, which is different from English. This suggests that the larger duration differences resulted from the difference in stress placement between Thai and English.

We then analyzed the effects of general English word classes, i.e., content or function words. Figure 9 shows the average of individual duration differences (the proposed prediction-error-based measure) observed in phones in content words subtracted by that observed in phones in function words. In the case of native English speakers, no negative values were found. Thus, this result indicates that native English speakers exhibited larger predicted duration differences for content words than for function words. In contrast, we found negative values in the case of Thai learners, with some strongly negative values compared with the median of the Thai learner group. These results suggest that some Thai learners used a different duration control mechanism for content and function words compared with native English speakers. Similarly to the case of English stress, we found that most learners who produced such duration differences had no or minimal experience of study in English-as-an-official-language countries. Therefore, the results suggest that these word classes can be considered as a key feature in evaluating the English skill of a Thai learner.

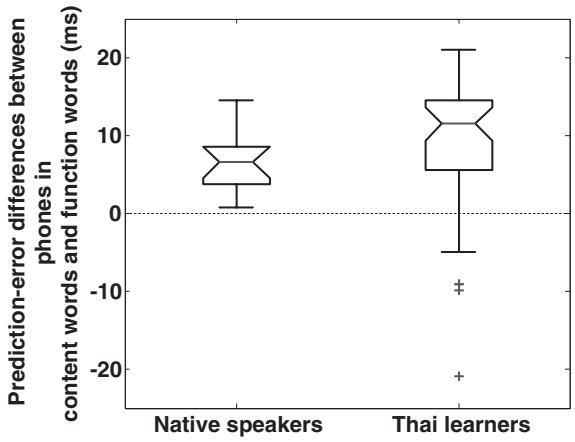

Fig. 9 Comparison between native speakers and learners for prediction-error differences of phone durations in content and function words. Each plot represents the average of individual prediction-errors in content words subtracted by that in function words.

5. CONCLUSIONS

We proposed a model-based automatic evaluation method and an objective measure for analyzing the speech-duration characteristics of Thai learners of English to evaluate the proficiency of English segmental duration control. The proposed method is based on an objective measure of actual segmental duration differences from durations predicted by a statistical duration model. An experiment was conducted to measure the duration differences of multiple groups of learners with different amount of English-study experience, and its results showed the effectiveness of the proposed objective evaluation method, in which statistical duration characteristics based on a generalized English duration model are used as a reference. Compared with other previously proposed measures, the proposed measure gave superior results for the discrimination between native English and Thai learner groups. Furthermore, the proposed measure of duration difference was also able to reveal the English segmental duration characteristics of Thai-native English learners. Our findings are promising for the quantitatively objective analysis of English skills.

ACKNOWLEDGEMENTS

We would like to thank the Human Language Technology Laboratory, National Electronics and Computer Technology Center (HLT, NECTEC, Thailand) for collecting the English speech data of Thai learners and native speakers. We are also grateful to Professor Helen Meng (Chinese University of Hong Kong) for providing the speech data of native English speakers from the CUHK Chinese Learners of English Speech Corpus (CUCHLOE). Furthermore, we would like to thank Professor Michiko Nakano and her students; Izabelle Grenon and her friend; and the members of Sagisaka laboratory for their collaboration in the subjective evaluation. This work was supported in part by the Waseda University RISE research project titled "Analysis and modeling of human mechanism in speech and language processing" and a Grant-in-Aid for Scientific Research B, No. 20300069, from JSPS.

REFERENCES

[1] Council of Europe, "Phonological Competence," in *Common European Framework of Reference for Languages* (Cambridge University Press, Cambridge, 2001), pp. 116–117.
[2] "Interagency Language Roundtable Language Skill Level Descriptions: Speaking," *Interagency Language Roundtable Language Skill Level Descriptions*, Interagency Language Roundtable (ILR) (1985).
[3] "ACTFL Proficiency Guidelines — Speaking," *ACTFL Proficiency Guidelines*, American Council on the Teaching of Foreign Languages (1999).
[4] Y. Tono, T. Kaneko, H. Isahara, T. Saiga, E. Izumi and M. Narita, "The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography," *Proc. Asian Association for Lexicography* (*ASIALEX*) *Biennial*, pp. 7–17 (2001).
[5] L. F. Bachman, *Fundamental Considerations in Language Testing* (Oxford University Press, Oxford, 1990).
[6] M. Canale and M. Swain, "Theoretical bases of communicative approaches to second language teaching and testing," *Appl. Linguist.*, **1**, 1–47 (1980).
[7] D. H. Hymes, "On communicative competence," in *Sociolinguistics: Selected Readings*, J. B. Pride and J. Holmes, Eds. (Penguin, Harmondsworth, 1972).
[8] I. I. Bejar, "A preliminary study of raters for the test of spoken English," Educational Testing Service (ETS), New Jersey, RR-85-5 (1985).
[9] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev and M. Weintraub, "Automatic evaluation and training in English pronunciation," *Proc. Int. Conf. Spoken Language Processing* (*ICSLP*), pp. 1185–1188 (1990).
[10] C. Cucchiarini, H. Strik and L. Boves, "Automatic assessment of foreign speakers' pronunciation of Dutch," *Proc. Eurospeech*, pp. 713–716 (1997).
[11] L. Neumeyer, H. Franco, V. Digalakis and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Commun.*, **30**, 83–93 (2000).
[12] C. Cucchiarini, H. Strik, D. Binnenpoorte and L. Boves, "Towards an automatic oral proficiency test for Dutch as a second language: Automatic pronunciation assessment in read and spontaneous speech," *Proc. Integrating Speech Technology in* (*Language*) *Learning* (*InSTIL*), pp. 18–25 (2000).
[13] C. Cucchiarini, H. Strik and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *J. Acoust. Soc. Am.*, **107**, 989–999 (2000).
[14] C. Cucchiarini, H. Strik and L. Boves, "Quantitative assessment of second language learners' fluency: an automatic approach," *Proc. Int. Conf. Spoken Language Processing* (*ICSLP*), pp. 2619–2623 (1998).
[15] C. Cucchiarini, H. Strik and L. Boves, "Using speech recognition technology to assess foreign speakers' pronunciation of Dutch," *Proc. New Sounds*, pp. 61–68 (1997).
[16] H. Strik, C. Cucchiarini and D. Binnenpoorte, "L2 pronunci-

ation quality in read and spontaneous speech," *Proc. Int. Conf. Spoken Language Processing* (*ICSLP*), pp. 582–585 (2000).

[17] N. Herry and D. Hirst, "Subjective and objective evaluation of the prosody of English spoken by French speakers: the contribution of computer assisted learning," *Proc. Speech Prosody*, pp. 383–386 (2002).

[18] J. Bernstein, J. De Jong and D. Pisoni, "Two experiments on automatic scoring of spoken language proficiency," *Proc. Integrating Speech Technology in* (*Language*) *Learning* (*InSTIL*), pp. 57–61 (2000).

[19] X. Xi, K. Zechner and I. I. Bejar, "Extracting meaningful speech features to support diagnostic feedback: An ECD approach to automated scoring," *Proc. Annu. Meet. Natl. Counc. Meas. Educ.* (*NCME*) (2006).

[20] X. Xi, D. Higgins, K. Zechner and D. M. Williamson, "Automated scoring of spontaneous speech using SpeechRater v1.0," Educational Testing Service (ETS), RR-08-62 (2008).

[21] K. Zechner and I. I. Bejar, "Towards automatic scoring of non-native spontaneous speech," *Proc. Human Language Technology* (*HLT*), pp. 216–223 (2006).

[22] K. Zechner, I. I. Bejar and R. Hemat, "Toward an understanding of the role of speech recognition in non-native speech assessment," Educational Testing Service (ETS), RR-07-02 (2007).

[23] "TOEFL iBT Scores: Better information about the ability to communicate in an academic setting," Educational Testing Service (ETS) (2005).

[24] C. Hayashi, "On the quantification of qualitative data from the mathematico-statistical point of view," *Ann. Inst. Stat. Math.*, **2**, 35–47 (1950).

[25] "The CMU Pronouncing Dictionary v.0.4," Carnegie Mellon University, Online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[26] M. P. Marcus, B. Santorini and M. A. Marcinkiewicz, "Building a large annotated corpus of English: the Penn treebank," *Comput. Linguist.*, **19**, 313–330 (1993).

[27] C. Hansakunbuntheung, H. Kato and Y. Sagisaka, "Model-based duration analysis on English natives and Thai learners," *Proc. ISCA Workshop on Experimental Linguistics* (*ExLing*), pp. 101–104 (2008).

[28] J. Kominek and A. W. Black, *CMU ARCTIC database for speech synthesis* (*version 0.95*) (2003).

[29] H. Meng, Y. Y. Lo, L. Wang and W. Y. Lau, "Deriving salient learners mispronunciations from cross-language phonological comparison," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding* (*ASRU*), pp. 437–442 (2007).

[30] L. Wang, X. Feng and H. Meng, "Mispronunciation detection based on cross-language phonological comparisons," *Proc. IEEE IET Int. Conf. Audio Lang. Image Process.* (*ICALIP*), pp. 301–311 (2008).

[31] "Acoustic model for adaptive ASR," VoxForge, Online: http://www.voxforge.org.

[32] A. Ito, T. Nagasawa, H. Ogasawara, M. Suzuki and S. Makino, "Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm," *Educ. Technol. Res.*, **29**, 13–23 (2006).

[33] Y. Yamashita, K. Kato and K. Nozawa, "Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison," *IEICE Trans. Inf. Syst.*, **E88-D**, 496–501 (2005).