INTERSPEECH 2009 BRIGHTON

Model-based automatic evaluation of L2 learner's English timing

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

[1] GITI / Language and Speech Science Research Labs, Waseda University, Tokyo, Japan
[2] NICT/ATR Media Information Science Laboratories, Kyoto, Japan

chatchawarnh@fuji.waseda.jp, kato@atr.jp, sagisaka@giti.waseda.ac.jp

10.21437/Interspeech.2009-42

Abstract

This paper proposes a method to automatically measure the timing characteristics of a second-language learner's speech as a means to evaluate language proficiency in speech production. We used the durational differences from native speakers' speech as an objective measure to evaluate the learner's timing characteristics. To provide flexible evaluation without the need to collect any additional English reference speech, we employed predicted segmental durations using a statistical duration model instead of measured raw durations of natives' speech. The proposed evaluation method was tested using English speech data uttered by Thai-native learners with different English-study experiences. An evaluation experiment shows that the proposed measure based on duration differences closely correlates to the subjects' English-study experiences. Moreover, segmental duration differences revealed Thai learners' speech-control characteristics in word-final stress assignment. These results support the effectiveness of the proposed model-based objective evaluation.

Index Terms: speech timing, quantitative evaluation, second language

1. Introduction

Spoken-language learning is a bi-directional 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 need some sorts of language proficiency measure to characterize their own current language proficiency levels, which allow them to monitor their further 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 proficiencies anytime without the need for a human rater.

The existing conventional language-proficiency evaluations, e.g. CEFR [1], ILR [2], ACTFL-OPI [3], TOEFL-iBT [4], generally provide some sort of subjective feedback by professional human raters. However, various problems of subjective evaluation have arisen and remain unresolved, such as the time-consuming nature of manual evaluation, inconsistency agreement among raters, raters' different and personal equations, and the need for multiple raters to reduce the raters' personal equations [5]. Therefore, automatic evaluation methods based on objective measures have been proposed to solve these problems.

Many research studies [6-13] have been conducted on the automatic evaluation of learner's proficiency. By using these kinds of automatic evaluation, interactive tests can be developed to provide immediate feedback to language learners. Nevertheless, these evaluations still do not clearly describe the precise quantitative factors that human raters use for evaluation. 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 to improving the tools and systems for language learning. Furthermore, this issue will particularly benefit self-learning language learners.

As emphasized in many frameworks for language-proficiency evaluation [1-4], timing is an essential issue. Furthermore, timing is a fundamental acoustic property of speech that can be directly measured from speech. Thus, this paper focuses on timing-based language learning evaluation. In this work, we propose an English-language duration model to measure the durational differences between a learner's segmental durations and those of natives. The use of a duration model enables a flexible choice of test sentences without needing 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 experiences of English study. In the following Section 2, we first introduce a proficiency evaluation using segmental duration differences and a statistical segmental duration model of native English. Next, in Section 3 and 4, we explain our experimental setup consisting of speech corpora, multiple groups of speakers with different English study experiences, and objective measurements of duration differences. In Section 5, experimental details and results are presented. Finally, in Section 6, the conclusions of this paper are given.

2. Model-based evaluation

To evaluate a learner's English proficiency based on timing control, we adopted an objective measure representing the average difference between segmental duration of a target learner and those of native speakers as an alternative to conventional subjective measures. To eliminate the need to collect additional comparable native-speaker reference speech data, a representative duration model of native English timing was statistically built using English speech data uttered by multiple native speakers. We measured the timing characteristics and duration differences between actual durations from speakers and statistically predicted ones from the model to compare the differences between natives and learners in English timing control.

For the computation of a statistical segmental duration of native English, we adopted segmental durations normalized by speech rate. Before modeling, we normalized segmental duration of each phone for each speaker using z-score normalization with mean and standard deviation (SD) to eliminate any speech-rate effect for inter-speaker 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 further used this mean as speaker-dependent speech rate for analysis. For the modeling, we adopted a multiple linear regression based on categorical factors [14] as shown in Eq. (1).

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

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

where N, $\hat{y}_i$, $\bar{y}$, x_{fc} and $\delta_{fc}(i)$ represent the number of data, the predicted duration of the i^{th} 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 i^{th} speech segment was encoded using the characteristic function $\delta_{fc}(i)$. By adopting the least-square-error minimization technique, modeling coefficients representing the contributions of the control factors were calculated.

As shown in Table 1, for control factors, we employed the current and four context phones, stress, phone position and the numbers of constituent phones in syllable, word and phrase, syllable position and the numbers of syllables in word and phrase, and the narrow and broad parts of speech. These factors were adopted by referring to the previous study on English duration [15].

Table 1 Control factors and categories employed in a linear regression modeling of normalized segmental duration of native English.

Factor	Category
Current phone	39 English phones [16]
Pre-preceding phone of current phone	39 English phones [16] and a pause
Preceding phone current phone	39 English phones [16] and a pause
Succeeding phone current phone	39 English phones [16] and a pause
Next succeeding phone of current phone	39 English phones [16] and a pause
Phone position in syllable	$P_{m,n}$; m = 1, …, n, n = 2, ..., 7
Numbers of constituent phones in syllable	1, …, 7
Phone position in word	See Fig. 4.
Numbers of constituent phones in word	1, …, 8, 9-10, 11-12, 13-14
Phone position in phrase	See Fig. 4.
Numbers of constituent phones in phrase	4-6, …, 73-75, 76-78
Syllabically lexical stress	Stressed, Unstressed
Syllable position in word	See Fig. 4.
Numbers of constituent syllables in word	1, …, 7
Syllable position in phrase	See Fig. 4.
Numbers of constituent syllables in phrase	1, …, 30
General part-of-speech	Function word, Content word
Specific part-of-speech	34 categories [17]

3. English speech corpora

We employed three types of English speech databases for evaluation. The first one was the ARCTIC database [18] read by English-speaking natives. This database was used for segmental duration modeling and to test this model's accuracy in reflecting native characteristics. The database was separated into two phonetically-balanced sets: ARCTIC set A with 593 sentences and ARCTIC set B with 539 sentences. The contents of these two sets were completely different. The second database was a read English speech database of the fairly tale "The north wind and the sun," from the CUCHLOE corpus [19], to test the proposed evaluation scheme. The sentences were uttered by English natives and speakers from English-as-an-official-language countries. The third database was also a test-speech database collected at NECTEC. It contained the same tale uttered by 45 Thai learners with different English-study background, and, one Indian English speaker.

The above databases established four groups used either for modeling or analysis. The first group consisted of ARCTIC set A uttered by four US speakers. It was used as the training set for the prediction of segmental durations by reference native speakers. The second one consisted of ARCTIC set B of the same four speakers. We referred to this group as a closed-speaker open-text set to evaluate the consistency of the model. If our evaluation scheme can be effectively used to calculate the duration differences between learners as a model-based approach, the predicted duration differences between the training and the open-text sets of the same speakers are expected to be closer than those of the learners' sets.

To evaluate the model's validity with various English accents, we used the third group as an open-speaker open-text set. It included three non-US-accent English speakers from ARCTIC set B, six speakers from CUCHLOE, and one speaker from NECTEC. The last group contains 45 Thai learners of English from NECTEC. We used this group as a test set to evaluate English duration characteristics of learners.

4. Objective duration-difference measures for proficiency evaluation

To evaluate the effectiveness of a measure using an English duration model, we compared the deviations of actual speech durations from the predicted durations for two speaker groups i.e. English natives and Thai learners. First, the speech data for evaluation were segmented by an HMM-based automatic segmentation scheme. We adopted HMM Toolkit (HTK) using adaptive acoustic model based on VoxForge speech database [20]. Then, we measured phone durations from the segmented speech data. Next, we used the English duration model with the control factors from the speech data to predict native English durations. Finally, we calculated root-mean-squared (RMS) differences between the measured and estimated reference duration.

5. Experimental results

5.1. Correlation between objective duration-difference measures and English study experience

Figure 1 shows a comparison of RMS duration differences from predicted durations between English native speakers and Thai learners. As the figure shows, the Thai-learner group produced the median of prediction difference of 58.1 ms

deviated from the English duration model, while the deviation of the English-native group is about 38.6 ms deviated from the model. Furthermore, the center quartiles of the distributions of English native 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.

To examine the relationship between the duration differences from predicted durations and English study experience more closely, we compared 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. As shown in Figure 2, noticeable duration differences were observed by learners' grouping according to the time spent in English education. The duration differences 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 consistency and reasonable prediction accuracy of the model both for the training and for the open set.

As for other-accented native speakers in the open-speaker data set, their duration differences were much closer to those of speakers in the model training 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 salient decrease in the distance from the reference model, while the learners with less experience in such countries showed larger duration differences with a large variation in English skills and wider phone duration variations than more experienced learners. In observing the correlation between period of time in the English-environment countries and duration differences, the results show a negative correlation coefficient of -0.37. These results support the effectiveness of the proposed objective measure. Since other out-of-scope background factors may also affect this measure, the wide variations of duration differences in Thai learner groups, especially in the least experienced group, can be found, and, need further investigation.

5.2. Duration-difference analysis to characterize English timing characteristics of Thai learners

To further investigate the effectiveness of the proposed duration differences, we analyzed duration differences between Thai learners and English native speakers. We observed the differences in 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 known as a stress-timed language like English, we analyze the duration differences caused by English stress by observing individual differences of average duration differences (the proposed prediction-error-based measure) of phones in stressed syllables subtracted by those of phones in unstressed syllables. In Figure 3, a number of the negative values of individual differences were mostly found in the learners' data. In case of English natives, we found this type of error in two of the non-US English natives. In case of Thai learners, most of the Thai learners who produced this type of error had no-or-least experience in English-as-an-official-language countries. This result suggests that a learner showing this negative value tends to use an odd type of control strategy comparing with native speakers to cope with stressing, and, a probable cause of the observed negative value is learner's 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 ends of a word or phrase than did the English native speakers. Figure 4 shows these characteristics 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, and, that stress factor shows possibility to be used as an evaluation factor.

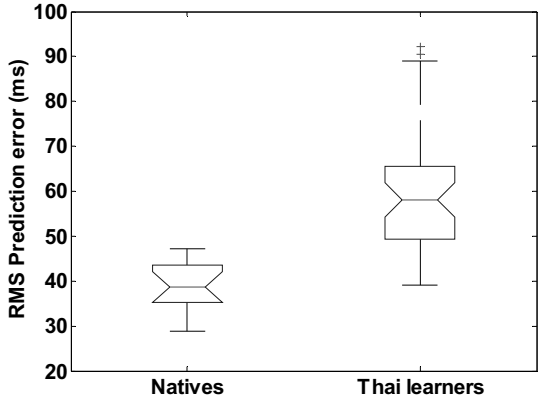

Figure 1 Comparison of RMS duration differences from predicted durations between English native speakers and Thai learners.

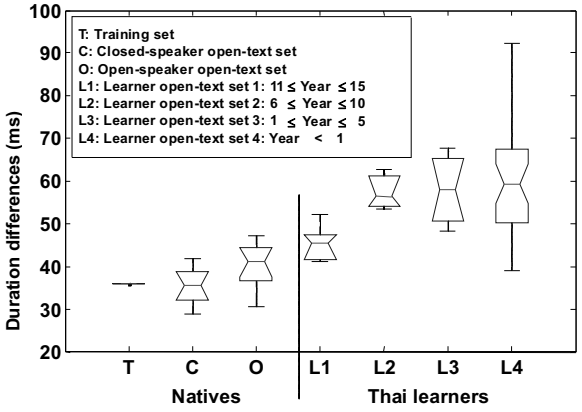

Figure 2 Comparison of RMS duration differences from predicted durations between English natives (C: closed speakers, O: open speakers) and Thai learners (L1 – L4), grouped by education period in English-as-an-official-language countries.

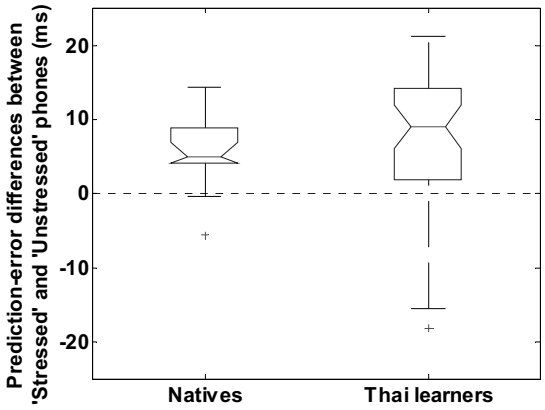

Figure 3 Comparison between natives and learners on prediction-error differences of phone durations at stressed and unstressed syllables.

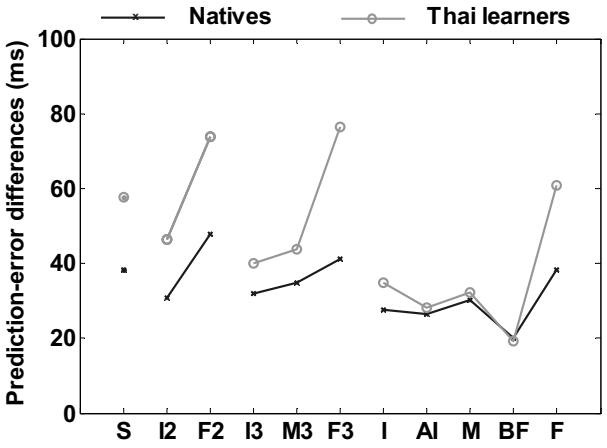

Figure 4 Duration differences from predicted durations at different word positions in different word-lengths between English natives and Thai learners (where "S", "I", "AI", "M", "BF", "F" labels represent syllable position in monosyllabic word, initial, after-initial, mid, before-final, and final position in a word, respectively. The indices of the positions represent word length. If not defined, they represent positions in multiple syllabic words with 4 syllables or more.).

6. Conclusions

We proposed a model-based automatic evaluation method and an objective measure to analyze the speech-duration characteristics of Thai learners of English for proficiency evaluation of English timing control. The proposed method is based on an objective measure of actual segmental duration differences from native-English durations predicted by a statistical duration model. An experiment was conducted to measure the duration differences of multiple groups of learners with different English-study experiences, and its results showed the effectiveness of the proposed objective evaluation method using statistical duration characteristics based on a generalized English duration model as a reference. Furthermore, the proposed duration difference measure was also able to reveal English timing characteristics of Thai-native English learners. Our findings are promising for making a quantitatively objective analysis of English skills.

7. 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 Prof. Helen Meng from Chinese University of Hong Kong (CUHK) for providing the English native speakers' speech data from the CUHK Chinese Learners of English Speech Corpus (CUCHLOE). This work was supported in part by the Waseda University RISE research project entitled "Analysis and modeling of human mechanism in speech and language processing" and a Grant-in-Aid for Scientific Research B, No. 20300069, of JSPS.

8. References

[1] Council of Europe, "Common European Framework of Reference for Languages", Online: http://www.coe.int/T/DG4/Linguistic/Source/Framework_EN.pdf , accessed on 24 Feb 2009, 116-117, 2001.

[2] Interagency Language Roundtable, "Interagency Language Roundtable Language Skill Level Descriptions: Speaking", Online: http://www.govtilr.org/Skills/ILRscale2.htm, accessed on 24 Feb 2009.

[3] American Council for the Teaching of Foreign Languages, "ACTFL Proficiency Guideline, ACTFL guidelines: Speaking", 1999.

[4] Educational Testing Service (ETS), "TOEFL iBT Scores: Better information about the ability to communicate in an academic setting", Online: http://www.ets.org/, accessed on 25 Feb 2009, 2005.

[5] Bejar, I., "A Preliminary Study of Raters for the Test of Spoken English", TOEFL Research Reports RR-85-5, Educational Testing Service (ETS), New Jersey, 1985.

[6] Bernstein, J., De Jong, J., Pisoni, D. and Townshend, B., "Two experiments on automatic scoring of spoken language proficiency", Proc. InSTIL2000 (P. Delcloque Ed.), 57-61, 2000.

[7] Cucchiarini, C., Strik, H. and Boves, L., "Quantitative assessment of second language learners'fluency by means of automatic speech recognition technology", J. Acoust. Soc. Am., 107 (2): 989-999, 2000.

[8] Cucchiarini, C., Strik, H. and Boves, L., "Using speech recognition technology to assess foreign speakers' pronunciation of Dutch", Proc. 3rd NEW SOUNDS, 61-67, 1997.

[9] Neumeyer, L., Franco, H., Digalakis, V. and Weintraub, M., "Automatic Scoring of Pronunciation Quality", J. Speech Communication, 30:83-93, 2000.

[10] Strik, H., Cucchiarini, C. and Binnenpoorte, D., "L2 Pronunciation Quality in Read and Spontaneous Speech", Proc. ICSLP-2000 and 6th ICSLP, 582-585, 2000.

[11] Zechner, K. and Xi, X., "Towards Automatic Scoring of a Test of Spoken Language with Heterogeneous Task Types", Proc. 3rd ACL-BEA 2008, 98-106, 2008.

[12] Xi, X., Zechner, K. and Bejar, I., "Extracting meaningful speech features to support diagnostic feedback: an ECD approach to automated scoring", NCME, 2006.

[13] Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R. and Butzberger, J. "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning". Proc. InSTiLL, 123-128, 2000.

[14] Hayashi, C., "On the Quantification of Qualitative Data from the Mathematic-Statistical Point of view", Annals of the Institute of Statistical Mathematics, Vol. 2, 1950.

[15] Hansakunbuntheung, C., Sagisaka, Y. and Kato, H., "Model-based duration analysis on English natives and Thai learners", Proc. ExLing, 101-104, 2008.

[16] The CMU Pronouncing Dictionary (version 0.4), Online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[17] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A., "Building a large annotated corpus of English: the Penn treebank," Computational Linguistics, Vol. 19, 313-330, 1993.

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

[19] Meng, H., Lo, Y.Y., Wang, L. and Lau, W.Y., "Deriving Salient Learners Mispronunciations From Cross-Language Phonological Comparison", Proc. ASRU, 2007.

[20] VoxForge's Acoustic model for adaptive ASR, Online: http://www.voxforge.org, accessed on 17 Apr 2008.