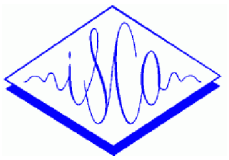

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Model-based Duration-difference Approach on Accent Evaluation of L2 Learner

Chatchawarn Hansakunbuntheung, Ananlada Chotimongkol,
Sumonmas Thatphithakkul, Patcharika Chootrakool

Speech and Audio Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand
{chatchawarn.han, ananlada.cho, sumonmas.tha, patcharika.cot}@nectec.or.th

10.21437/Interspeech.2012-224

Abstract

This paper aims at using a model-based duration-difference approach to analyze L2 learners' duration-aspect accent, and, segmental duration characteristics. We use the durational differences deviated from native-English speech duration as an objective measure to evaluate the learner's timing characteristics. The use of model-based approach provides flexible evaluation without the need to collect any additional English reference speech. The proposed evaluation method was tested on English speech data uttered by native English speakers and Thai-native English learners with different English-study experiences. The experimental results show speaker clusters grouped by English accents and L2 learners' English-study experiences. These results support the effectiveness of the proposed model-based objective evaluation.

Index Terms: speech timing, quantitative evaluation, second language

1. Introduction

Learning to speak a new language requires both speaking training and assessment in order to increase spoken proficiency. For beginners, a learning technique by listening and copying what they heard is a simple and general method used in language courses and computer-assisted language learning (CALL) systems. From the viewpoint of self-learning second language (L2) learners, learners need some sorts of spoken proficiency assessment to automatically characterize their own current spoken proficiency levels, which allow them to monitor their further progress. If an automatic assessment can be provided, L2 learners can evaluate themselves and keep track of their proficiencies anytime without the need for a human rater. However, proficiency scores alone do not provide sufficient feedback for L2 learners to pinpoint their speaking flaws. We need more informative feedback that can identify a L2 learner's weak points in speaking. Due to variation of English accents, some learners might find difficulty in distinguishing different English accents. Unintentionally, L2 learners might mix different accents together during listening and copying process. To train speaking accent close to the target native accent, spoken proficiency assessment should be able to suggest how learner's accent sound like the target-native accent.

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.

Thus, a model-based duration-difference approach [15], a quantitative language-proficiency evaluation, was introduced to fulfill the requirement of informative feedback by measuring the durational differences between a learner's segmental durations and ones predicted by a native English duration model. 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. The model-based method presented rather high correlation between its proposed scores and spoken proficiency scores rated by English instructors. It was also able to reveal the English segmental duration characteristics of Thai-native English learners comparing to the English natives. By observing the duration-difference score calculated from the model-based approach, the measurement showed different duration-difference scores among the speakers with different English background such as accents, experiences of English usage. This finding shows potential to further use the model-based duration-difference approach to discriminate L2 learner's accent and, also, to explore spoken duration characteristic difference of English accents.

Thus, this paper adopts the model-based duration-difference approach to aim at duration-aspect accent analysis and L2 learners' segmental duration characteristic analysis in more details. To evaluate the approach, 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 introduce a proficiency evaluation using segmental duration differences and a statistical segmental duration model of native English. Next, in Section 3, 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 4 and 5, experimental details and results are presented, recpectively. Finally, in Section 6, the conclusions of this paper are given.

2. Model-based timing 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. Fig.1 shows an overview of the model-based duration-difference approach. 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.

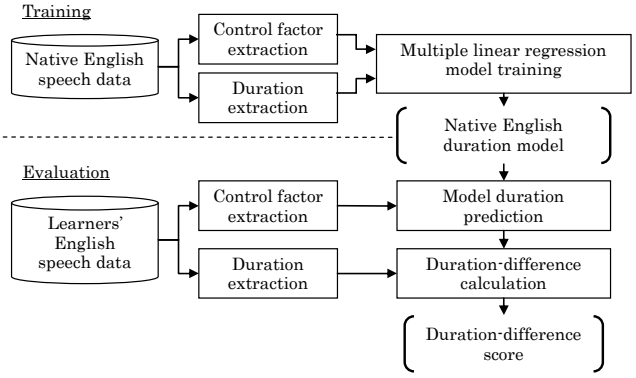

Figure 1 Overview of Model-based duration-difference Evaluation approach

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], pause
Preceding phone of current phone	39 English phones [16], pause
Succeeding phone of current phone	39 English phones [16], pause
Next succeeding phone of current phone	39 English phones [16], pause
Phone position in syllable	$P_{m,n}$; m = 1, …, n, n = 2, ..., 7
Phone position in word	I2, F2, I3, M3, F3, I, AI, M, BF, F, S
Phone position in phrase	
Syllable position in word	
Syllable position in phrase	
No. of constituent phones in syllable	1, …, 7
No. of constituent phones in word	1, …, 8, 9-10, 11-12, 13-14
No. of constituent phones in phrase	4-6, …, 73-75, 76-78
No. of constituent syllables in word	1, …, 7
No. of constituent syllables in phrase	1, …, 30
Syllabically lexical stress	Stressed, Unstressed
General part-of-speech	Function word, Content word
Specific part-of-speech	34 categories [17]

3. Speech material

We employed three types of English speech databases for evaluation as shown in the data summary in Table 2. The first one was the ARCTIC database [18] read by English-speaking natives. 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. Set A database was used for modeling a standard duration model that represents native English segmental duration characteristics. Set B was used for testing the accuracy of the model in reflecting native characteristics. The second database was a read English speech database of the fable "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 fable uttered by 45 Thai learners with different English-study background, and, one Indian English speaker.

The above databases established four groups that were 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.

Table 2 Summary of English speech data of native English and L2 speakers.

	Speech data set			
	Training	Test set 1	Test set 2	Learner set
Speech data source	CMU ARCTIC set A	CMU ARCTIC set B	CMU ARCTIC set B, CUCHLOE, NECTEC	NECTEC
Speaker set	Closed	Closed	Open	Open
Text set	Closed	Open	Open	Open
Number of speakers	4	4	10	45
Native language	US English	US English	US and Non-US English	Thai

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. Evaluation

4.1. Objective duration-difference measures for proficiency evaluation

To compare overall proficiency scores among speakers, an overall duration-difference score is calculated for each speaker. An overall score represents duration deviations of a learner's segmental duration characteristics comparing to the referenced ones predicted by the native English duration model. The learner's segmental duration characteristics are a set of actual speech durations collected from read speech data mentioned in Section 3. To collect a learner's duration data set automatically, 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, phone durations from the segmented speech data were measured. Next, we used the English duration model with the control factors from the same speech data to predict native English duration data set. Finally, we calculated root-mean-squared (RMS) differences between the measured and estimated reference duration.

4.2. Speaker-Feature duration-difference matrix for multidimensional scaling analysis

To observe a learner's specific duration characteristic of a duration control factor mentioned in Table 1, a duration-difference score of a specific control factor category, feature-based score, is calculated for each category. For a speaker, a set of feature-based scores of all control factor categories in Table 1 are collected as a speaker vector as shown in Fig. 2. Then, a collection matrix of all speaker vectors, called "Speaker-Feature duration difference matrix", as shown in Fig. 2 is constructed for multidimensional scaling (MDS) analysis.

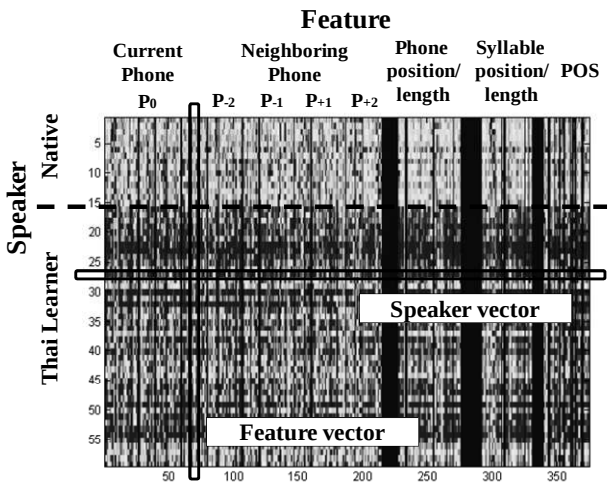

Figure 2 Speaker-Feature duration-difference matrix.

5. Experimental results

Using the duration difference measure mentioned in Section 4.1, a relationship between the duration difference and speakers' English was found as shown in Fig. 3 [15]. A correlation of -0.62 between phone duration-difference scores and subjective score on Thai-native English learners' speech data was also found. When closely observing the groups of Thai learners grouping by study period in English-as-an-official-language country, 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 phone duration differences than more experienced learners.

To explore the relationship between speakers and feature-based scores using MDS analysis, we performed the analysis by varying its dimension from 2 to 5. By considering the Scree plot using the elbow criterion, we got a suitable dimension of 3 with the stress of 0.082. Fig. 4 present the result of the MDS analysis based on the model-based duration-difference approach. It shows the groups of speakers grouped by speakers' accent and English experience. Considering the grouping of speakers, it clearly shows the positions of all English natives are on the left-hand side of the plot, while the positions of the learners are on the right-hand side. Another arrangement of the plot found here is that the groups of non-US natives and Thai natives who have education period in English-as-an-official-language countries tend to locate close to the group of US natives on the left-hand side. In addition, we can notice that the positions of the speakers with the same English accents tend to be close together. For example, the positions of US natives, which are the group of the target accent, are grouped together on the most left-hand side. We also found a group of Hong Kong native, who speak English as an official language, is located closely on the right of the US native group. Other English native speakers with various accents seem to distribute on the left side but not tight as other native English with the same accents. Considering the groups of Thai-native English learners with different education background in English-speaking countries, groups of learners with more education period in the English speaking countries tend to locate tighter and closer to the left side of the plot than the less ones. The result in Fig. 4 shows correspondence to Fig. 3, with more grouping details.

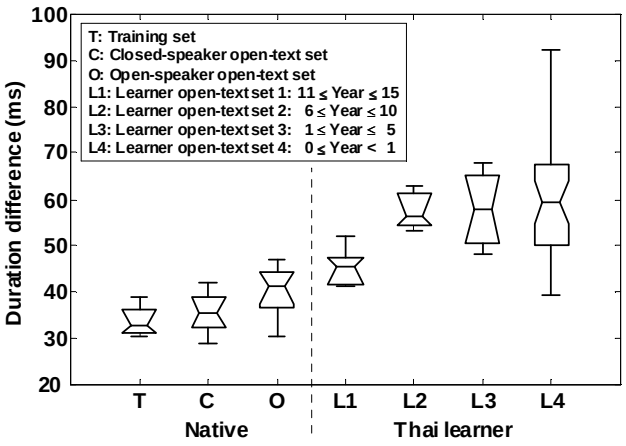

Figure 3 Comparison of RMS duration differences from predicted durations between English natives (C: closed speakers, O: open speakers) and Thai learners (L1 – L4), grouped by education period in English-as-an-official-language countries.

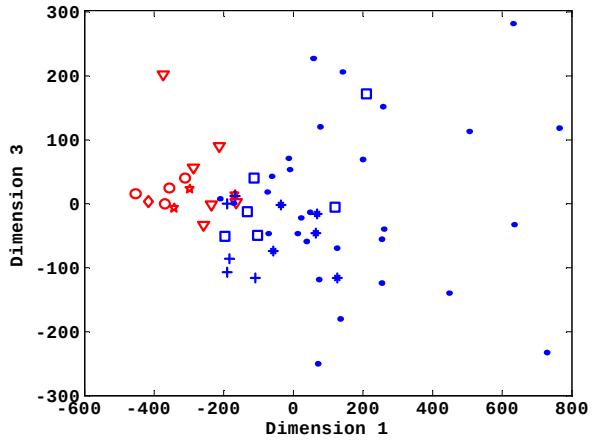

Figure 4 MDS analysis of speaker-feature duration difference matrix (markers: '○' US native, '◇' UK native, '☆' Hong Kong native, '△' other English native, '+' Thai learners group L1, '*' Thai learners group L2, '□' Thai learners group L3, '.' Thai learners group L4).

6. Conclusions

This paper adopted the model-based duration-difference approach to make a duration-aspect accent analysis of L2 learners. The proposed method is based on an objective measure of actual segmental duration differences from durations predicted by a statistical duration model. The experiment was tested on English speech data uttered by various native English speakers and Thai-native learners with different English-study experiences. An experimental result showed the grouping of speakers by English accents and also found the grouping of speakers by the L2 learners' English-study experiences in English-as-an-official-language countries. Conclusively, the experiment results show the potential use of the model-based duration-difference approach for L2 learner's accent assessment and spoken proficiency evaluation.

7. Acknowledgements

We would like to thank Prof. Yoshinori Sagisaka, Prof. Nakano from Language and Speech Science Research Labs at Waseda University for many kind supports and evaluation. 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).

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/, 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 automatic evaluation of second-language learner's English segmental duration characteristics", Acoust. Sci. & Tech., 31(4):267-277, 2010.

[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.