EVALUATION OF JAPANESE MANNERS OF GENERATING WORD ACCENT OF ENGLISH BASED ON A STRESSED SYLLABLE DETECTION TECHNIQUE

Yukiko FUJISAWA, Nobuaki MINEMATSU, and Seiichi NAKAGAWA
{*fuji, mine, nakagawa*}*@slp.tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Technology,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441-8580 JAPAN

ABSTRACT

Different languages sometimes use different acoustic manners to transmit the same kind of linguistic information. This fact lets us easily suppose that learners tend to transmit the information in a manner of not a language to learn but their native language. While English word accent is linguistically almost the same as Japanese one, the word accent acoustically differs between the two languages. This paper aims at automating the detection of the generated word accent and the evaluation of how it is generated. By using context-sensitive HMMs, stressed and unstressed syllables were modeled separately for their structure and for their position in a word. Results showed that 90% of the stressed syllables were correctly detected in open experiments. In the matching process, each of the likelihood scores derived from different parameters was multiplied with its weighting factor. The optimal combination of the factors for the detection can be thought to reflect each speaker's own manner of the accent generation. Differences of the quasi-optimal combinations between Japanese and native speakers mainly accorded with findings reported in previous studies on English teaching to foreign learners.

1. INTRODUCTION

Recent advances of speech recognition techniques provide us with non-native speech processing as one of major research challenges. Unlike native speech, the distribution of acoustic features of the non-native speech is considered broadly distorted. And the distortion depends on learners' native languages, their dialects, the degree of learning how to pronounce words and sentences of the target language, and so forth. Namely, the distortion should be found to be dependent on each learner. This indicates that, especially with the aim of instructing a learner, it is important to find wrong habits in *his* pronunciation, report them, and instruct the learner how to correct them. Although several previous studies proposed methods to deal with non-native speech for its evaluation, most of them only tried to detect pronunciation errors in the utterance and make the scoring scheme more correlated with the human scheme[1]-[4]. Namely, what were discussed there are how to *detect* the errors and how to *score* the utterance, not to *instruct* the learner how to correct the errors.

It can be found between two languages that even the same kind of linguistic information is transmitted in different acoustic manners. And it is easily assumed that the learners tend to transmit the information in a manner of not a language to learn but their native language, and that this is a main cause of the language dependent habit in the learners' pronunciation. While an acoustic event called 'word accent' is said to have almost the same linguistic role between English and Japanese, its acoustic realization differs between them. Japanese word accent is represented by an F_0 contour of the word and English one is characterized by power, duration, F_0, and vowel quality. According to a previous study in phonetics, the above assumption is valid as for the word accent[5]. And it is interesting that English words with wrong accents are reported to be more difficult for native speakers to accept than those with wrong phonemic features so long as the wrong features are produced consistently[6][7]. These findings indicate the importance for the learners to learn to generate the accent with correct acoustic features and in a correct position.

On these backgrounds, a method for stressed syllable detection was proposed in our previous studies[8][9]. In this paper, two experiments are carried out. One is for the improvement of the detection method and the other is for the automatic estimation of Japanese manners of generating the word accent of English, which is aiming at evaluating English words spoken by Japanese. What is estimated here can also be used for the automatic diagnosis and instruction of the learners.

2. AUTOMATIC DETECTION OF STRESSED SYLLABLES

2.1. Modeling (Un)stressed Syllables

In English phonology, *word accent* is often used to indicate *word stress*. And a stress label is usually assigned to a syllable, not to a phoneme. Before describing the modeling methods of the stress, it is beneficial to review the structure of syllables of English and Japanese and the difference between them. While almost all the Japanese syllables consist of CV or V, English ones have more various forms. According to [10], an English syllable has a central vowel and sequences of consonants located before and after the vowel. Length of the preceding sequence is from zero to three and that of the succeeding one is from zero to four. It follows that the longest syllable of English is CCCVCCCC. This structural difference surely leads to the difference in the number of kinds of syllables. While there are only one hundred or so kinds in Japanese, English is estimated to have more than ten thousand varieties[11].

In this paper, to model (un)stressed syllables, 1) 1 to 4 dimensions of LPC mel cepstrum coefficients and their derivatives, 2) power and its derivative, 3) $F0$ and its derivative were used to make a parameter vector[12]. Using this parameterization of speech signals, continuous density HMMs with duration control were utilized. In this modeling, the Viterbi score at time t and state i is calculated as

$$f(i,t) = \max_{j,\tau} \left[f(j, t-\tau) a_{ji} d_i(\tau)^{\phi} \prod_{k=1}^{\tau} b_i(y_{t+1-k}) \right], \quad (1)$$

where a_{ji}, $d_i(\tau)$, and $b_i(y_t)$ indicate a transition probability, a duration probability, and an output probability density function respectively. And ϕ is a weighting factor for $d_i(\tau)$. Assuming no correlation between any two of the above three acoustic parameters, $b_i(y_t)$ can be written as

$$b_i(y_t) = \prod_{s=1}^{3} b_i^s(y_t^s)^{\rho_s}, \qquad \text{where} \quad \sum_{s=1}^{3} \rho_s = 3. \quad (2)$$

Here, y_t^s represents a sub-vector corresponding to one of the above three acoustic parameters and ρ_s is a weighting factor for $b_i^s(y_t^s)$. In the experiments described in this section, all the weighting factors were set to be 1.0.

If the number of kinds of syllables in English is limited as in Japanese, the acoustic models can be built for individual syllables. A large variety in English syllables, however, required us to make the models separately for each syllable *class*. And the class-based modeling led us to use only a small number of dimensions of cepstrum coefficients, which was four in this study. And the following five schemes for the syllable clustering were examined taking *structural* or *positional* information of the syllable into account.

a) 2 classes; stressed and unstressed syllables. This is the simplest clustering.

b) 6 classes; S_H, S_T, and S_O separately for stressed and unstressed syllables, where S_H/S_T denote syllables at the head/tail of a word and S_O indicates a syllable at the other parts of the word. Here, *positional* information of a syllable in the word is integrated into HMMs.

An observed $F0$ contour shows a rising pattern at the beginning of an utterance and a falling pattern at the end, which is *language independent*. And this is the case even when the utterance is an isolated word[13]. It means that the first and the last syllables in a word should be separately modeled at least in terms of its $F0$ contour.

c) 16 classes; V_S, CV_S, V_SC, CV_SC, V_L, CV_L, V_LC, and CV_LC separately for stressed and unstressed syllables, where V_S/V_L denote short/long vowels and C denotes a sequence of consonants. In this case, *structural* information of a syllable is introduced into the HMMs.

d) 48 classes; both of the two kinds of information are considered. Namely, syllables are modeled separately for their structure and their position in a word.

e) 80 classes; context information is introduced to refine the clustering **d**. Here, 4 kinds of labels – stressed/unstressed/head_of_word/tail_of_word – are used as left and right context information.

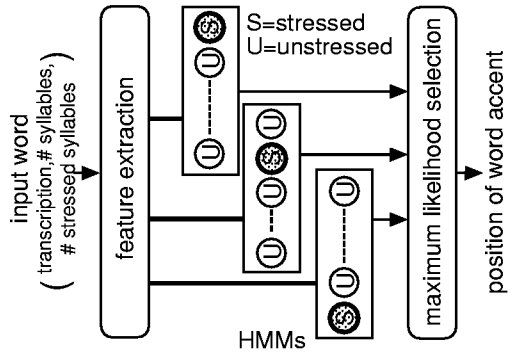

Figure 1: Stressed syllable detection experiment

Prosodic features are also called 'supra-segmental' features. It implies that the distribution of acoustic features of a syllable varies dependently on those of surrounding syllables.

Words with more than one syllable were extracted from ATR English word database (2 male British speakers; spk-**1**&**2**) and used as training data. Then, syllable models were built separately for each speaker using approximately 3,300 words. Through all the experiments, speech materials were digitized with 12 kHz and 16 bit sampling and the acoustic analysis was performed using 21.3 msec frame length and 8.0 msec frame rate. $F0$ and power were also extracted with the same rate and, after being transformed to a logarithmic scale, they were normalized to have zero as the averaged values over each utterance. When building the models, $F0$ values for unvoiced segments were required. For these segments, $F0$ values were estimated by linear interpolation of the preceding/succeeding voiced segments.

2.2. Detection Experiments

Detection of a stressed syllable in an input word was carried out based on the maximum likelihood criterion using a word-level score. An input word was matched with a concatenation of stressed/unstressed HMMs. Here, a syllabic transcription of the word, the number of syllables and that of stressed syllables of the word (one in this experiment) were all treated as *given*. Hence, the number of candidate stress patterns was N for an input word with N syllables. Position of the stressed HMM in the concatenation which produces the highest word-level score was identified as a *stressed* syllable, which is shown in Figure **1**.

Table **1** shows the correct detection rates for all the conditions **a** to **e** for each training speaker. The above database was also used as testing data. Namely, both of closed and open experiments were carried out. In the closed experiments, it can be seen that more informative models give us higher rates. And the highest averaged rate for the open experiments is also found in the case **e** (context-sensitive), which shows the validity of the proposed finer modeling.

Table 1: Results of stressed syllable detection [%]

train \ test	spk-**1**					spk-**2**				
	a	**b**	**c**	**d**	**e**	**a**	**b**	**c**	**d**	**e**
spk-**1**	77.4	88.9	86.4	93.9	95.8	79.9	79.9	88.2	88.9	91.1
spk-**2**	74.1	74.1	84.5	90.6	89.7	77.4	90.1	88.7	93.8	95.8

3. EVALUATION OF JAPANESE MANNERS OF GENERATING WORD ACCENT

3.1. Proposed Method for the Evaluation

As mentioned in Section **2.**, the Viterbi score in the matching process for the stressed syllable detection is calculated as equation (**1**) and the output probability density function can be written as equation (**2**). Then, we can get a resulting formula for the Viterbi score as

$$f(i,t) = \max_{j,\tau} \left[f(j,t-\tau) a_{ji} d_i(\tau)^{\phi} \prod_{k=1}^{\tau} \prod_{s=1}^{3} b_i^s(y_{t+1-k}^s)^{\rho_s} \right].$$

This equation can be interpreted as producing the Viterbi score $f(i,t)$ by multiplying sub-scores $d_i(\tau)$ and $b_i^s(y_t^s)$ with weighting factors ϕ and ρ_s. Namely, the score is obtained by integrating the observed distributions of acoustic features on tempo($d_i(\tau)$), spectrum($b_i^1(y_t^1)$), power($b_i^2(y_t^2)$), and tone($b_i^3(y_t^3)$) with adequate weighting factors.

In the stressed syllable detection experiments, all the weighing factors, ϕ and ρ_s, were set to be 1.0 in the training and the testing. However, it can be easily assumed that the combination of the factors (ρ_1,ρ_2,ρ_3,ϕ) which gives us the highest detection rate is *not* (1.0,1.0,1.0,1.0) even if the (un)stressed models are trained with all the factors being 1.0. This should be especially the case when the testing data is spoken by non-native speakers because of the acoustic distortion produced by the speakers. In this section, the combination which gives the highest rate (henceforth, the optimal combination) is estimated. Increase of a factor for an acoustic feature mainly indicates that more focus should be placed on the feature for more correct detection. Namely, modifying the factors can be considered changing 'hearing characteristics' of computers and the optimal combination is thought to reflect the acoustic features dominantly used for accent generation, in other words, the pronunciation habits in learners' utterances.

Several experiments were carried out to verify the proposed method in the following sections. Considering some findings reported in previous studies on English teaching and the authors' subjective views on English pronunciation by Japanese, the following differences in the optimal combination were expected to be found between Japanese learners and native speakers.

- ρ_1(native) > ρ_1(Japanese) [spectrum]
- ρ_2(native) > ρ_2(Japanese) [power]
- ρ_3(native) < ρ_3(Japanese) [tone]
- ϕ(native) > ϕ(Japanese) [tempo]

3.2. Estimation of the Optimal Combination and Its Analysis

Since the aim of this paper is to verify the above method, the estimation of the *strictly* optimal combination was not done. Instead, the best combination among a given set of combinations, that is the *quasi*-optimal combination, was calculated and its differences between Japanese and native speakers were analyzed. To obtain the combination, the following procedures were carried out, where, in addition to the ATR database, a few hundreds of word utterances were used as testing data listed in **Table 2**.

Table 2: Testing data used in the experiments
N/J represent Native and Japanese. B/A/C represent British, American, and Canadian respectively.

speaker	N/J	#words
spk–**1**	N(B)	3334
spk–**2**	N(B)	3290
spk–**3**	N(A)	60
spk–**4**	N(C)	45

speaker	N/J	#words
spk–**5**	J	48
spk–**6**	J	53
spk–**7**	J	51
spk–**8**	J	58
spk–**9**	J	44

i) Firstly, the factors are set to be (1.0,1.0,1.0,ϕ). Then, by modifying the ϕ, the model configurations – speaker (**1**/**2**) and format (**a** to **d**) – which intermediately provide the first and the second best detection rates are determined separately for each testing speaker. Here, ϕ can be 0,1,2,3,..9,10,15, or 20. And Φ_1 and Φ_2 will be used to refer to the first and the second best ϕs.

ii) Next, the factors are set to be (ρ_1,ρ_2,ρ_3,Φ_i). Then, by modifying the ρ_i, the quasi-optimal combinations, which finally give the first and the second best detection rates, are searched for. In this step, the following combinations are examined as ρ_i.

$$(\rho_1,\rho_2,\rho_3) \in \begin{cases} (1.0,1.0,1.0) \\ (1.5,0.5,1.0),(1.5,1.0,0.5),(0.5,1.5,1.0) \\ (1.0,1.5,0.5),(0.5,1.0,1.5),(1.0,0.5,1.5) \\ (2.0,0.0,1.0),(2.0,1.0,0.0),(0.0,2.0,1.0) \\ (1.0,2.0,0.0),(0.0,1.0,2.0),(1.0,0.0,2.0) \\ (2.0,0.5,0.5),(0.5,2.0,0.5),(0.5,0.5,2.0) \end{cases}$$

It should be noted that all the materials spoken by Japanese in **Table 2** satisfy the following conditions. 1) two English teachers perceive the stressed syllable in the same position of the word. 2) they acknowledge that the word accent is intelligible enough.

The quasi-optimal combinations of each of Japanese and native speakers are shown with their highest detection rates in **Tables 3** and **4** respectively. In the tables, the model format **e** is *not* used as mentioned above, which is currently investigated. And in **Table 3**, only the open experiments are carried out for spk-**1** and spk-**2**.

Table 3 shows that the integration of the *structural* information, i.e., model formats **c** and **d**, works effectively in native utterances. In **Table 4**, however, 3 subjects out of 5 have their quasi-optimal combinations in other formats than **c** or **d**. This indicates that Japanese learners tend to utter English words with inappropriate syllable structure, which is well-known as a fact that Japanese are apt to insert an additional vowel between successive consonants. According to statistical analysis of variance, ρ_1 of native speakers is significantly larger than that of Japanese ($p < 0.01$). Also as for ρ_3, significant difference was found between Japanese and native speakers (ρ_3(N) < ρ_3(J); $p < 0.01$). This means that Japanese learners are inclined to generate English word accent mainly by manipulating

Table 3: Quasi-optimal combinations for native speakers

testing speaker	format	training speaker	weighting factors (ρ_1,ρ_2,ρ_3,ϕ)			rate [%]
spk-**1**	**d**	**2**	(0.5, 2.0, 0.5,	7∼9	)	92.7
spk-**2**	**d**	**1**	(0.5, 1.0, 1.5,	5/6	)	90.9
spk-**3**	**d**	**1**	(1.0, 1.5, 0.5,	15/20	)	93.3
			(0.5, 2.0, 0.5,	15	)	
	b	**2**	(0.5, 2.0, 0.5,	2	)	
	a	**2**	(0.5, 2.0, 0.5,	1∼3	)	
spk-**4**	**c**	**1**	(1.0, 1.0, 1.0,	1∼3	)	86.7
			(0.5, 1.5, 1.0,	9/10	)	
			(1.5, 0.5, 1.0,	1∼4	)	
			(1.5, 1.0, 0.5,	1∼4	)	
			(0.5, 2.0, 0.5,	1/2/9/10	)	
	a	**2**	(1.0, 1.0, 1.0,	9/10	)	
			(0.5, 2.0, 0.5,	5/9/10/20	)	
			(1.0, 1.5, 0.5,	9/10	)	

Table 4: Quasi-optimal combinations for Japanese

testing speaker	format	training speaker	weighting factors (ρ_1,ρ_2,ρ_3,ϕ)			rate [%]
spk-**5**	**b**	**1**	(0.5, 1.0, 1.5,	1∼3	)	85.4
			(0.5, 1.5, 1.0,	1/2	)	
			(0.5, 2.0, 0.5,	1	)	
	c	**1**	(0.0, 2.0, 1.0,	6∼9	)	
			(0.0, 1.0, 2.0,	7∼9	)	
			(0.5, 0.5, 2.0,	7∼9	)	
spk-**6**	**b**	**2**	(0.5, 1.0, 1.5,	3∼10	)	88.7
			(1.5, 0.5, 1.0,	7	)	
			(2.0, 0.0, 1.0,	9/10	)	
spk-**7**	**b**	**2**	(1.0, 1.0, 1.0,	5∼10	)	98.0
			(0.5, 1.5, 1.0,	5∼10	)	
			(0.5, 1.0, 1.5,	5∼8	)	
			(1.0, 0.5, 1.5,	6∼8	)	
spk-**8**	**c**	**1**	(0.5, 1.5, 1.0,	10/15/20	)	100
	c	**2**	(0.5, 1.0, 1.5,	15	)	
spk-**9**	**a**	**2**	(0.5, 2.0, 0.5,	1∼10	)	100
			(0.0, 2.0, 1.0,	1∼10	)	
			(0.5, 1.5, 1.0,	1∼7	)	
			(0.5, 1.0, 1.5,	6∼10	)	
			(0.5, 0.5, 2.0,	5∼10	)	

tone of speech, which is exactly a manner of generating word accent of Japanese. On the other hand, no difference in ϕ was unexpectedly found even at the significance level of 0.1. This result is considered due to the difference among the processing schemes of power, F_0, and duration. Unlike the first two parameters, which were normalized to have zero as the averaged values over each utterance, the duration was utilized without normalization. It may be necessary to introduce the normalization process using the averaged syllable length to the duration control of HMMs.

4. CONCLUSIONS

In order to develop a method evaluating Japanese manners of generating English word accent, two experiments were carried out in this study. One was for detecting a syllable where a learner located the word accent and the other was for estimating and evaluating the learner's manner of generating the word accent. By modeling (un)stressed syllables using context-sensitive HMMs, 96% and 90% of the stressed syllables were correctly detected in speaker-dependent and speaker-independent experiments respectively. The learner's manner of generating English word accent was estimated by searching for the quasi-optimal combinations of weighting factors for the four acoustic features used for calculating the scores in the stressed syllable detection. Differences of the combinations between Japanese and native speakers mainly accorded with findings reported in previous studies on English teaching, which indicates the validity of our proposed method. As future works, in addition to the detection and evaluation experiments using a larger mount of speech material, we are planning more exact calculation of the optimal combination, introduction of the normalized duration control into the HMMs, visualization of the optimal combination, automatic diagnosis, generation of the effective instructions, and their feedback to the learners.

REFERENCES

1. S. Hiller *et al.*, "SPELL: An automated system for computer-aided pronunciation teaching," *Speech Communication 13,* pp.463-473 (1993).
2. H. Hamada *et al.*, "Automatic evaluation of English pronunciation based on speech recognition techniques," *IEICE Trans. vol. E76-D, No.3,* pp.352-359 (1993).
3. Y. Kim *et al.*, "Automatic pronunciation scoring of specific phone segments for language instruction," *Proc. EUROSPEECH'97,* pp.645-648 (1997).
4. H. Franco *et al.*, "Automatic pronunciation scoring for language instruction," *Proc. ICASSP'97,* pp.1471-1474 (1997).
5. Y. Shibuya, "Differences between native and non-native speakers' realization of stress-related durational patterns in American English," *J. Acoust. Soc. Am., Vol. 100, No.4, Pt.2,* pp.2725 (1996).
6. G. Kawai *et al.*, "An experimental study on the reliability of scoring pronunciation of English spoken by Japanese students," *Technical Report of IEICE, ET95-44,* pp.89-96 (1995, in Japanese).
7. A. Cutler *et al.*, "The predominance of strong initial syllable in the English vocabulary," *Computer Speech and Language, 2,* pp.133-142 (1987).
8. N. Minematsu *et al.*, "Automatic detection of accent in English words spoken by Japanese students," *Proc. EUROSPEECH'97,* pp.701-704 (1997).
9. Y. Fujisawa *et al.*, "Detection of stressed syllables in English words using HMMs and acoustic evaluation of the realized stress," *Technical Report of IEICE, SP98-46,* pp.7-14 (1998, in Japanese).
10. S. Takebayashi *et al.*, "A primer of English phonetics (shokyū eigo onsei gaku)," published by Taishūkan shoten (1991, in Japanese).
11. L. Rabiner *et al.*, "Fundamentals of speech recognition," *Prentice-Hall,* p.436 (1993).
12. A. Ljolje *et al.*, "Recognition of isolated prosodic patterns using Hidden Markov Models," *Computer Speech and Language,* 2, pp.27-33(1987).
13. H. Fujisaki *et al.*, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J.Acoust. Soc. Jpn., 4,* pp.233-242, (1984, in Japanese).