Instantaneous Estimation of Accentuation Habits for Japanese Students to Learn English Pronunciation

Naoki NAKAMURA† *Nobuaki MINEMATSU*‡ *Seiichi NAKAGAWA*†

†Toyohashi University of Technology ‡ University of Tokyo,

naoki@slp.ics.tut.ac.jp mine@gavo.t.u-tokyo.ac.jp nakagawa@slp.ics.tut.ac.jp

Abstract

More and more efforts have been recently made to apply speech technologies to language learning[1]–[4]. The authors have been especially focusing on Japanese manners of generating English word stress. This is because accentuation habits inevitable to Japanese learners can be easily found in their stress generation. In our previous studies, a stressed syllable detector and an accentuation habit estimator were developed[5]–[7], where the estimated habits of individual learners accorded well with their English pronunciation proficiency rated by English teachers. However, the estimation methods in our previous studies required several dozens of word utterances or a relatively large amount of computation even when a single word utterance enabled the estimation. In this paper, we investigated a method which required only a single word utterance with a small computation cost. Results showed that similar tendencies can be found between the habits estimated in our previous study and those in the current one.

1. Introduction

Rapid internationalization imposes two different tasks on speech engineers. One task is developing domain-independent speech-to-speech interpreters, using which humans are allowed to speak only their mother tongues. The other task is applying speech technologies to assisting humans' second or third language learning, where they can learn the language effectively and efficiently. Comparison between the ability of computers and that of humans to process spoken languages allows us to suppose that completing the latter is more practical and realistic.

In applying speech technologies to assisting language learning, it is very important to consider characteristics of the native language of the learner and those of the target language. As is well known, English and Japanese are quite different linguistically and phonetically. And we can easily find pronunciation habits in English spoken by Japanese. One typical example is word accent. Although word accent is *linguistically* almost the same between Japanese and English, it *acoustically* differs between them. This phenomenon causes an accentuation habit inevitable to Japanese learners. Since Japanese word accent is characterized by an $\boldsymbol{F_0}$ contour of the word, Japanese learners tend to generate English word accent mainly by manipulating $\boldsymbol{F_0}$[8], although it should be generated by controlling four acoustic factors of vowel quality, power, $\boldsymbol{F_0}$ and duration[1].

In our previous study, a method of estimating the accentuation habit was proposed[6], where the habit was defined as acoustic features dominantly used for accent generation and they were estimated by using an HMM-based stressed syllable detector[5]. And a method of visualizing the estimated habit was also proposed[6], where the *abstract* and *integrated* representation of the above four factors was realized. Experiments showed that the visualized habits corresponded well to pronunciation proficiency scores of individual learners rated by four English teachers.

In [6], however, the estimation could be done only after a learner pronounced several dozens of words because the habit estimator utilized stressed syllable detection *rates*. Although a method was proposed to estimate the habit with a single word utterance[7], it required a relatively large amount of computation. In this paper, we propose a new method which uses ratios of likelihood sub-scores and enables the habit estimation by using only a single word utterance with a small computation cost. By adopting the proposed method, feedback on the habit can be provided for a learner *interactively*, which should surely motivate him or her for learning further *continuously*.

2. Automatic estimation of the accentuation habit

2.1. Modeling (un)stressed syllables[5]

Speech samples were digitized with 12 kHz and 16 bit sampling. The 14-th order LPC analysis was carried out using 21.3 msec window length and 8.0 msec frame rate. $\boldsymbol{F_0}$ and power were also extracted with the same rate and, after being transformed to logarithmic scale, they were normalized to have zero as mean values over each sample. The following three streams were used to make a parameter vector; 1) the first four ones of LPC mel cepstrum coefficients and their Δs, 2) power and its Δ, and 3) $\boldsymbol{F_0}$ and its Δ. Using this parameterization, Continuous Den-

sity HMMs (CDHMMs) with duration control were built assuming no correlation between any two of the above three streams. In this study, English syllables were classified into 48 syllable groups in terms of their accentual, positional and structural attributes. And each of the groups was modeled by the above CDHMMs.

2.2. Detection of stressed syllables[5]

Using the syllable group HMMs, a stressed syllable detector was implemented based upon the maximum likelihood criterion using a word-level score. An input word was matched with candidate stress patterns. A candidate stress pattern was formed as a concatenation of a stressed HMM and unstressed ones. In the detection, a syllabic transcription of the word, the number of syllables and that of stressed syllables of the word (one in this study) were all treated as given. The position of the stressed HMM in the concatenation which produced the highest word-level score was identified as *stressed*.

2.3. Estimation of the accentuation habit[6]

In the matching procedure, the Viterbi score at time t and state i is calculated as

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

where a_{ji}, $d_i(\tau)$ and $b_i^s(y_t^s)$ indicate a transition probability, a duration probability, and an output probability density of a sub-vector respectively. A sub-vector y_t^s indicates one of cepstrum-, power-, and pitch-related parameters. And ϕ and ρ_s are weights of $d_i(\tau)$ and $b_i^s(y_t^s)$. This equation can be interpreted such that the score is obtained by integrating the sub-scores of the observed acoustic features on vowel quality ($b_i^1(y_t^1)$), power ($b_i^2(y_t^2)$), pitch ($b_i^3(y_t^3)$) and duration ($d_i(\tau)$) with their weights ρ_s and ϕ.

In the training phase, all the weights were fixed to be 1.0. However, this weight combination is easily supposed *not* to be the optimal combination for stress detection especially for non-native learners. Increase of a weight in the detection phase is mainly interpreted to emphasize its corresponding feature. Therefore, the optimal combination is thought to reflect the acoustic feature dominantly used for stress generation, which is called the accentuation habit of individual learners in the current study.

2.4. Visualization of the Estimated Habit[6]

The optimal combination was decided out of a prepared set of weight combinations. For duration weight (ϕ), it was varied from 0.0 to 20.0 with a step of 0.5, which gave us 41 variations. As for the other weights (ρ_s), they were varied satisfying a condition $\sum_s \rho_s = 3.0$ ($\rho_s \geq 0$). In other words, $\{\rho_s\}$ were prepared so that they were distributed evenly on

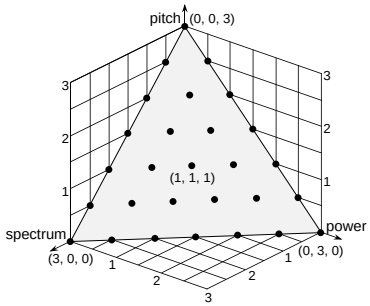

Figure 1: Distribution of weight combinations of $\{\rho_s\}$

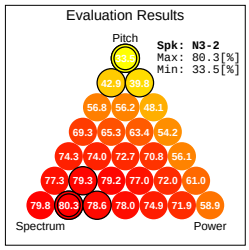

Figure 2: The representative triangle of the habit

a triangle, a pitch/spectrum/power triangle, shown inFigure 1. The number of combinations is 28. The visualization of the pronunciation habit could be obtained by representing detection rates at individual dots by different colors. But since a triangle could be drawn per duration weight, a learner came to get as many as 41 triangles. So the representative triangle was derived as an expected pattern of the 41 triangles along with a duration weight axis. An example of the resulting representatives is shown in Figure 2, where two double circles indicate the maximum and the minimum of detection rates, henceforth the maximum/minimum circle. In our previous work[6], the high accordance was observed between the visualized habits of individual learners and their English pronunciation proficiency. In this method, however, the estimation could be done only after a learner pronounced tens of words. To make up this defect, a method of *instantaneous* estimation requiring only a *single* word utterance is devised below.

3. Instantaneous and rapid estimation of the accentuation habit

3.1. Instantaneous estimation of the habit[7]

In our previous study[7], a method for the instantaneous estimation was proposed, where likelihood ratios at each weight combination were calculated as

$$\log R(w) = \log L(w|\lambda_c) - \max_{j \neq c} \log L(w|\lambda_j). \quad (1)$$

Here, w is an input word. And λ_j and λ_c indicate HMMs for stress pattern j and those for the intended (*correct*) pattern respectively. Using $\log \mathcal{R}(w)$ at each dot, a triangle was drawn and the weight combination which maximized the difference between the likelihood score of the correct pattern and the high-

est score of the other competing patterns was treated as the accentuation habit. A representative pattern of the 41 triangles was simply defined as an average pattern along a duration weight axis. The total number of likelihood calculations of HMMs is $28 \times 41 \times$(number of syllables in word).

3.2. Instantaneous and rapid estimation

In the previous section, the habit was estimated and visualized by comparing likelihood ratios at every weight combination. Therefore, it was required to calculate a likelihood score a large number of times. To reduce the computation cost, in this paper, we use a ratio of likelihood sub-score ratios of each feature.

Firstly, the likelihood sub-score of feature s is calculated as

$$\log L_s = \sum_V \log b_i^s(y_t^s)^{\rho_s}, \quad s = 1, 2, \text{ and } 3, \qquad (2)$$

where V is the Viterbi path obtained in Section 2.3 and all the weights of $\{\rho_s\}$ are fixed to 1.0. Secondly, for feature s, the likelihood sub-score ratio, R_s, is calculated in the following equation.

$$\log R_s(w) = \log L_s(w|\lambda_c) - L_s(w|\lambda_J) \qquad (3)$$

λ_J is the stress pattern which maximizes the total score of the competing patterns in Equation (1) in the case of $\{\rho_s\} = (1.0, 1.0, 1.0)$. And the maximum of the three R_ss is multiplied by n for emphasizing the accentuation habit. In this study, $n = 2$ is experimentally used. Finally, the ratio of likelihood sub-score ratios of each feature is calculated as

$$Q_{\text{spe}} = \frac{\log R_1}{\sum_s \log R_s}, \; Q_{\text{pow}} = \frac{\log R_2}{\sum_s \log R_s}, \; Q_{\text{pit}} = \frac{\log R_3}{\sum_s \log R_s}. \qquad (4)$$

$\log R_s$ sometimes turns out to be negative. In these cases, $\log R_s$ is fixed to be 0.0. By using Q values, the accentuation habit is drawn almost in the same manner as in the previous studies. The total number of likelihood calculations is only "number of syllables in word".

3.3. Evaluation of the proposed method

To evaluate the proposed method, several experiments were designed and carried out. The stressed HMMs and the unstressed ones were built for each syllable group using speech samples of Table 1, which are a part of ATR English word database. And using the HMMs, *whole* pronunciation habits and *partial* habits of individual speakers of set **A** and **J** were estimated and visualized. Here, the *whole* habits are those estimated by using multiple word utterances in the previous study while the *partial* habits are those by using a single utterance. Set **A** are a part of RM1 isolated word database and set **J** are speech samples recorded in a sound-proof room in our laboratory.

Table 1: Speech samples for training HMMs

set	#spk	native lang.	vocab. size	#words
B	1	British	3,334	3,334

Table 2: Speech samples for evaluating the method

set	#spk	native lang.	vocab. size	#words
A	7	American	381	546
J	7	Japanese	60	341

Figure 3 shows the locations of the maximum and the minimum neighboring circles for each of seven Japanese and seven Americans (*whole* habits, see Section 2.4)[6]. Figure 4 shows the locations of the highest likelihood ratios of each word utterance (*partial* habits, see Section 3.1)[7]. Size of circle indicates the frequency of the highest ratio being observed at the location. Figure 5 shows the locations of the estimated habits for each word utterance in the current study (*new partial* habits, see Section 3.2).

As for Figure 4, Japanese habits are frequently distributed close to or at the pitch vertex, while Americans ones are often distributed around or at the spectrum vertex. It means that these distributions of the *partial* habits are similar to those of the *whole* habits of Figure 3, which clearly indicates the validity of the method proposed in [7]. To investigate the distributions of habits of Figures 4 and 5, we divide a triangle into 4 parts(spectrum, power, pitch and middle part). Figure 6 shows the distribution of habits of Japanese and Americans (*partial* habits, see Section 3.1). Figure 7 shows the distribution of habits (*new partial* habits, see Section 3.2). About a half of Japanese habits are distributed over the pitch part (54.0% and 51.1%), while American habits are distributed over the spectrum part with larger possibility than that of any other part (34.8% and 47.4%). These distributions are similar to the distribution of Figure 3, and it indicates the validity of the method proposed in this paper.

4. Conclusion

In this paper, a new method to estimate the accentuation habit with a small computation cost was proposed, where only a single word utterance was required. The reduction of computation was realized by using ratios of likelihood sub-score ratios.

5. References

[1] S. Hiller *et al.*, "SPELL: An automated system for computer-aided pronunciation teaching," Speech Communication, vol.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] C. Cucchiarini *et al.*, "Quantitative assessment of second language learners' fluency: An automatic approach," Proc. ICSLP'98, vol.6, pp.2619–2622 (1988).

[4] R. Akabane-Yamada *et al.*, "Computer-based second language production training by using spectro-

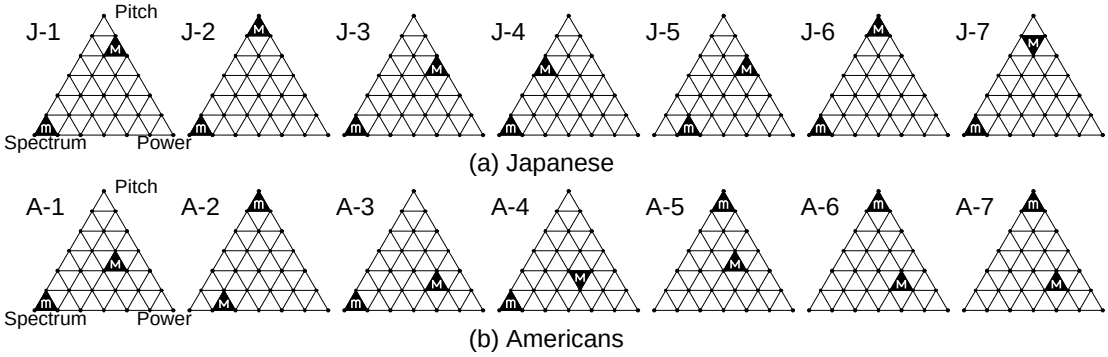

Figure 3: Whole habits of Japanese (upper) and Americans (lower) estimated in the previous study

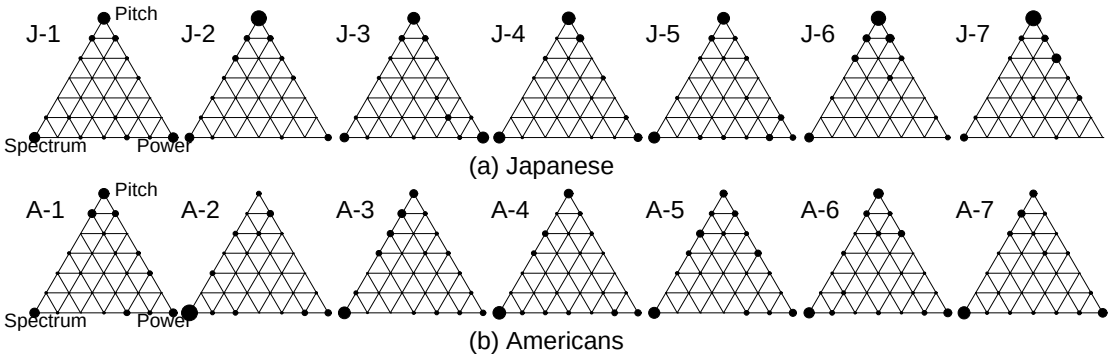

Figure 4: Partial habits of Japanese (upper) and Americans (lower) estimated in the previous study

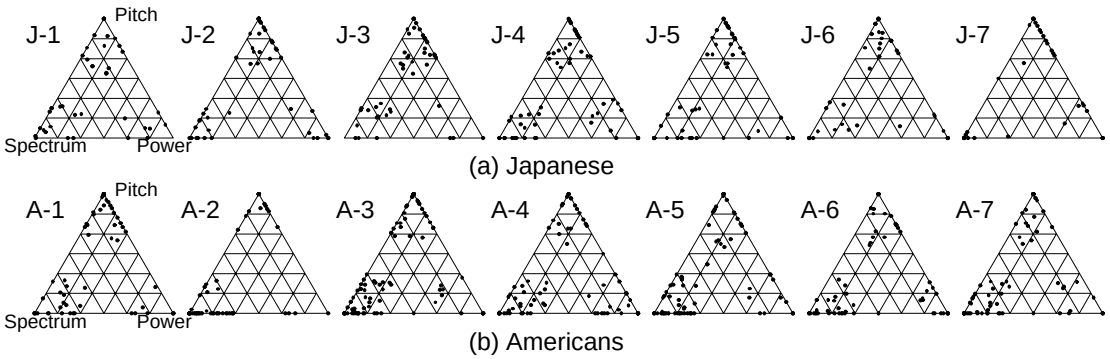

Figure 5: Partial habits of Japanese (upper) and Americans (lower) estimated in the proposed method

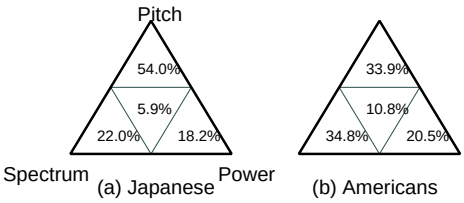

Figure 6: Distribution of partial habits of Japanese (left) and Americans (right)

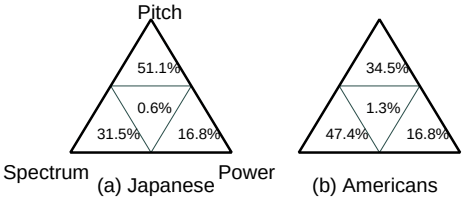

Figure 7: Distribution of new partial habits of Japanese (left) and Americans (right)

graphic representation and HMM-based speech recognition scores," Proc. ICSLP'98, vol.5, pp.1747–1750 (1998).

[5] N. Minematsu *et al.*, "Automatic detection of accent in English words spoken by Japanese students," Proc. EUROSPEECH'97, pp.701–704 (1997).

[6] N. Minematsu *et al.*, "Visualization of pronunciation habits based upon abstract representation of acoustic observations," Proc. INSTIL'2000, pp.130–137 (2000).

[7] N. Minematsu *et al.*, "Instantaneous estimation of prosodic pronunciation habits for interactive instructions to Japanese learning English," Proc. ICSLP'2000, vol.3, pp.191–194 (2000).

[8] 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).