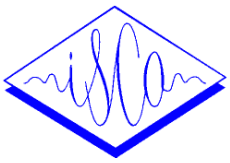

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

INSTANTANEOUS ESTIMATION OF PROSODIC PRONUNCIATION HABITS FOR JAPANESE STUDENTS TO LEARN ENGLISH PRONUNCIATION

Nobuaki MINEMATSU† and *Seiichi NAKAGAWA*‡
mine@gavo.t.u-tokyo.ac.jp *nakagawa@slp.ics.tut.ac.jp*

† Graduate School of Engineering, University of Tokyo
‡ Department of Information and Computer Sciences, Toyohashi University of Technology

ABSTRACT

More and more efforts have been recently made to apply speech technologies to language learning and develop CALL systems[1]–[4]. The authors have been focusing on Japanese manners of generating English word stress. This is because pronunciation habits which are inevitable to Japanese learners can be easily found in the stress generation. In our previous study, a stressed syllable detector and a pronunciation habit estimator were developed[5], where the estimated habits of individual learners accorded well with their English pronunciation proficiency rated by English teachers. However, the habit could be estimated only after a learner pronounced several dozens of words because the habit estimator referred to stressed syllable detection *rates*. In this paper, using another criterion, a method of instantaneous estimation was proposed which required only a *single* word utterance. Results showed that an average pattern of the instantaneously estimated habits accorded well with the habits obtained in our previous study.

1. INTRODUCTION

Rapid internationalization imposes two different tasks on speech engineers. One task is the development of domain-independent speech-to-speech interpreters, using which humans are allowed to speak only their mother tongues. The other one is the application of 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 a pronunciation habit inevitable to Japanese learners. Since Japanese word accent is characterized by an F_0 contour of the word, Japanese learners tend to generate English word accent mainly by manipulating F_0[6], although it should be generated by controlling four factors of vowel quality, power, F_0 and duration[1].

In our previous study, a method of estimating the pronunciation habit was proposed[7], 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[8]. And a method of visualizing the estimated habit was also proposed[9], 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.

However, the estimation could be done only after a learner pronounced several dozens of words because the habit estimator utilized stressed syllable detection *rates*. Here, various combinations of weights were prepared for calculating sub-scores, each of which was related to one of the four factors. The total score was obtained by multiplying the weighted four sub-scores. And the weight combination maximizing the detection rate was supposed to reflect acoustic features dominantly used for accent generation. In this paper, using another criterion, a method of *instantaneous* estimation is proposed which requires only a *single* word utterance. In the new criterion, the weight combination which maximizes the difference between the matching score of the correct stress pattern and the highest matching score of the other competing patterns is treated as the habit. By using this 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 PRONUNCIATION HABIT

2.1. Modeling (Un)stressed syllables

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. 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) F_0 and its Δ. Using this parameterization, 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[8]

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 au-

tomatically segmented into syllables, and then they were matched with their corresponding HMMs in the candidate stress patterns. A 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 Pronunciation Habit[7]

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

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

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 $b_i^s(y_t^s)$ respectively. 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 1.0. However, this weight combination is easily supposed *not* to be the optimal combination for stress detection especially for non-native learners' utterances. 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 pronunciation habit of individual learners in the current study.

2.4. Visualization of the Estimated Habit[9]

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 a triangle, a pitch/spectrum/power triangle, shown in **Figure 1**. 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, and two sets of three single circles mean the maximum and the minimum of *averaged* detection rates over *neighboring* three circles, henceforth the maximum/minimum neighboring circles. In our previous work[9], 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.

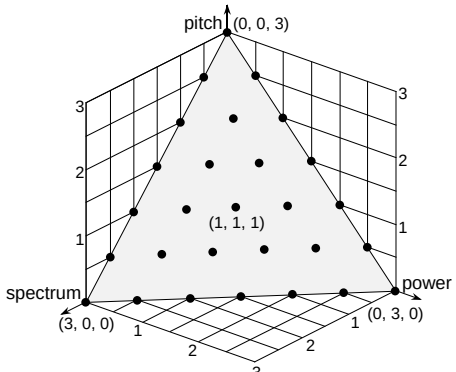

Figure 1: Distribution of weight combinations of $\{\rho_s\}$

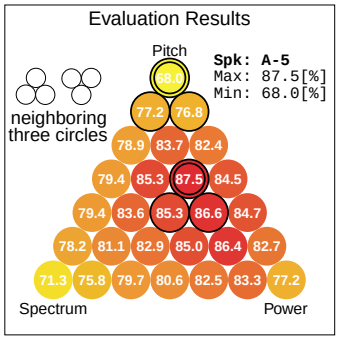

Figure 2: The representative triangle of the habit

3. INSTANTANEOUS ESTIMATION OF THE PRONUNCIATION HABIT

3.1. Conditions for the Estimation

It can be easily supposed that learners are highly motivated for continuous learning by the instantaneous habit estimation and its interactive feedback. Assuming that the estimated habit obtained in our previous study[9] is correct as *whole* habit of the learner, the instantaneously estimated habit, which is referred to as *partial* habit in the rest of the paper, should satisfy the following conditions.

1. The estimation and visualization procedures must be done separately for each word utterance.
2. An average pattern of a learner's *partial* habits accords well with his or her *whole* habit.

The first condition requires us not to use the detection *rates* to make a triangle. In this study, a method using only likelihood ratios at individual dots (weight combinations) on a triangle, which can be obtained per word, is proposed.

3.2. Instantaneous Estimation of the Habit

Since our objective is to estimate the pronunciation habit by using only a single word utterance, the following learning situation can be considered; a learner is provided on a monitor with what word to pronounce and which syllable to stress. In this situation, a transcription of the word, the

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

number of the syllables and the intended position of the stress are all treated as known to a habit estimator.

A triangle can surely be drawn for each utterance by using likelihood scores of the 28 weight combinations at a duration weight. In this case, however, comparison between any two scores of the 28 combinations makes no sense because scores at different dots use different equations for calculating the likelihood scores (see Equation (**1**)). To make the comparison between two dots meaningful, we used likelihood ratio at each dot which is calculated below.

$$\log \mathcal{R}(w) = \log \mathcal{L}(w|\lambda_c) - \max_{j \neq c} \log \mathcal{L}(w|\lambda_j) \qquad (2)$$

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 highest score of the other competing patterns was treated as the pronunciation habit. A representative pattern of the 41 triangles was simply defined as an average pattern along a duration weight axis.

Although the above procedure could give us a single representative triangle for each word, a problem remained to be solved as for averaging representative triangles of different words. The range of likelihood ratio of a word was quite different from another. In the worst case, an average of representative triangles of a word set was almost the same as a representative one of a specific word in the set. To solve this problem, the linear normalization was introduced before averaging the representatives. Here, the maximum and the minimum of likelihood ratios in a representative triangle were converted to 100.0 and 0.0 respectively.

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, the whole pronunciation habits and the partial habits of individual speakers of set **A** and **J** were estimated and visualized. Set **A** are a part of RM1 isolated word database and set **J** are speech samples which were recorded in a sound-proof room in our laboratory. Using these habits, the evaluation was done based upon condition 2 in Section **3.1.**.

Figure 3 show the locations of the maximum and the minimum neighboring circles for each of seven Americans (**A**). **Figure 4** show the locations of these circles of the normalized and averaged likelihood ratio. The former figures are drawn by using the *whole* habits and the latter are done by using average patterns of the *partial* ones. Figures **5** and **6** do similarly for each of seven Japanese learners (**J**). As is reported in our previous study[5], the locations of the maximum neighboring circles of the whole habits are closer to the pitch (top) vertex in Japanese than in Americans. And as for Japanese, the distance from the pitch vertex to the maximum neighboring circles is larger with speakers of higher English pronunciation proficiency.

Except for a few speakers or a few maximum/minimum neighboring circles, the relatively good accordance is seen between the whole habits and the averaged partial habits. The correlation between expected detection rates of the whole habits and normalized and averaged likelihood ratios of the partial habits was calculated. And it was 0.72 and 0.83 on the average for Americans and Japanese respectively. In the figures of the partial habits, the correlation is shown for each speaker. Although two speakers, **A-4** and **J-5**, showed quite low values, all the other speakers gave us rather high correlation. It indicates that the proposed method almost satisfies condition 2 in Section **3.1.**. The observed discordances are considered partly due to speech samples whose likelihood ratios are less than 0.0 at every dot in a triangle. Since the stress detection with these samples fails at any dot, they are probably illegal utterances and should have been rejected out of the experiments.

Results of further analysis are shown in **Figures 7** and **8**, which indicate the probabilities that the maximum/minimum neighboring circles of likelihood ratio, namely, the partial habit, are found on each location on a triangle. They are represented in the form of the percentage. It should be noted that the probabilities in the figure are calculated by using all the partial habits of word samples, not by using the normalized and averaged habits shown in **Figures 4** and **6**. Clearly shown here, the maximum/minimum neighboring circles tend to be found on edges of the triangles, although the maximum neighboring circles are often found inside the triangle in the case of Americans' normalized and averaged partial habits in **Figure 4**. It means that using only one feature rather than multiple ones can give us clearer separation of the correct pattern's likelihood and the other patterns' likelihood. And shown in **Figures 7** and **8**, the probability of each feature (vertex) showing the maximum/minimum likelihood ratio is quite different between Japanese and Americans. Based upon this finding, we are currently developing a simplified scheme of estimating a partial habit by using three features separately[10].

4. CONCLUSIONS

In this paper, a method of instantaneously estimating the pronunciation habit was proposed by modifying a habit estimation technique previously devised by the authors. Since the devised technique had used stressed syllable detection rates, it could not estimate the habit word by word. The proposed technique in this study utilized likelihood ratios, not scores, on the triangular representation to make the estimation process instantaneous. Evaluation experiments showed that average patterns of the instantaneously estimated habits accorded well with the whole pronunci-

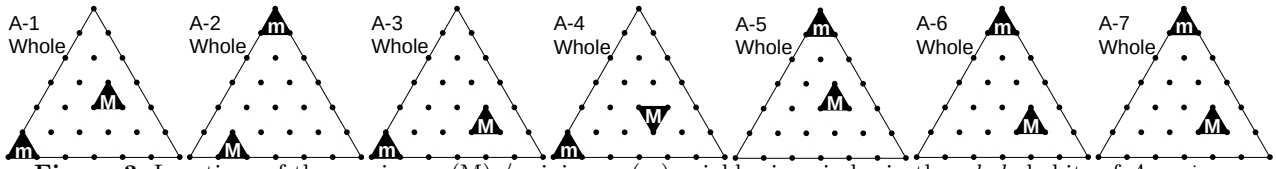

Figure 3: Locations of the maximum (M) / minimum (m) neighboring circles in the *whole* habits of *Americans*

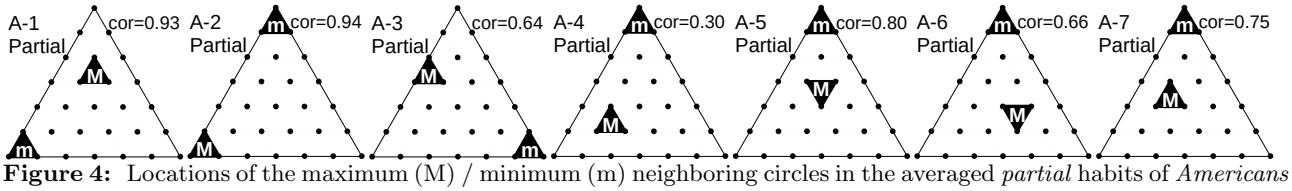

Figure 4: Locations of the maximum (M) / minimum (m) neighboring circles in the averaged *partial* habits of *Americans*

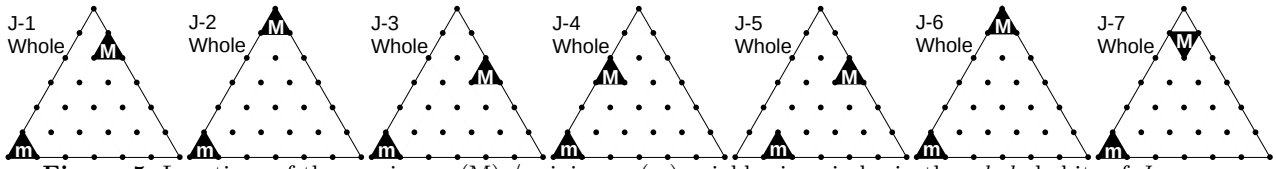

Figure 5: Locations of the maximum (M) / minimum (m) neighboring circles in the *whole* habits of *Japanese*

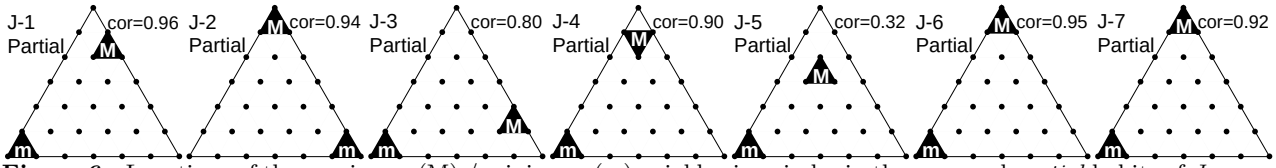

Figure 6: Locations of the maximum (M) / minimum (m) neighboring circles in the averaged *partial* habits of *Japanese*

ation habits obtained in our previous study. And further analysis indicated that the maximum/minimum likelihood ratios are likely to be obtained by using only one feature rather than multiple ones in matching process. This implies that the instantaneous estimation can be further simplified. In addition to implementing the simplification, the evaluation of the proposed method as a CAI tool in a classroom is left as one of future works.

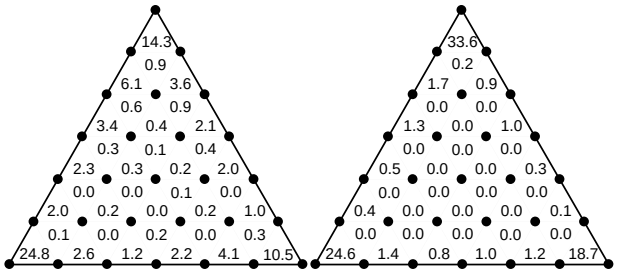

Figure 7: Probabilities of the maximum (left-hand) or the minimum (right-hand) being found on each location for Americans in the form of percentage

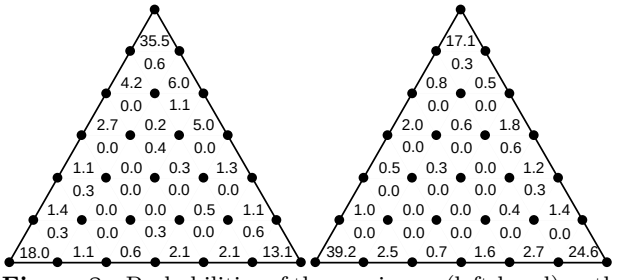

Figure 8: Probabilities of the maximum (left-hand) or the minimum (right-hand) being found on each location for Japanese in the form of percentage

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-graphic representation and HMM-based speech recognition scores," Proc. ICSLP'98, vol.5, pp.1747–1750 (1998).
5. N. Minematsu *et al.*, "Estimation of pronunciation habits and its application to prosodic evaluation of pronunciation proficiency," Proc. ICCE'99, vol.2, pp.177–200 (1999).
6. 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).
7. Y. Fujisawa *et al.*, "Evaluation of Japanese manners of generating English word accent based on a stressed syllable detection technique," Proc. ICSLP'98, pp.3103–3106 (1998).
8. N. Minematsu *et al.*, "Automatic detection of accent in English words spoken by Japanese students," Proc. EUROSPEECH'97, pp.701–704 (1997).
9. N. Minematsu *et al.*, "Prosodic evaluation of English words spoken by Japanese based upon estimating their pronunciation habits," Proc. ICSP'99, pp.439–444 (1999).
10. N. Minematsu *et al.*, "Automatic estimation of pronunciation habits using a single word utterance based upon a stressed syllable detection technique," Technical Report of IEICE, SP2000-2, pp.9-16 (2000, in Japanese).