The Seventh Western Pacific Regional Acoustics Conference
WESTPRAC VII
Kumamoto, Japan, 3-5 October 2000

VISUALIZATION OF PRONUNCIATION HABITS USING SPEECH RECOGNITION TECHNIQUES

Nobuaki MINEMATSU† and Seiichi NAKAGAWA‡

† University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN
E-mail: mine@gavo.t.u-tokyo.ac.jp

‡ Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580 JAPAN
E-mail: nakagawa@ics.tut.ac.jp

ABSTRACT

Different languages sometimes use different acoustic manners to transmit the same kind of linguistic information. This fact lets us easily suppose that language 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 phenomenon causes pronunciation habits which are *inevitable* and *peculiar* to Japanese learners of English. In this paper, two issues are described; 1) the automatic estimation of the prosodic pronunciation habit and 2) the visualization of the estimated habit to provide the learners with easy-to-understand feedback instructions. In the first issue, the habit was defined as a acoustic feature dominantly used for the word stress generation. And the habit was estimated by using a stressed syllable detector proposed by the authors previously. And in the second issue, the habit, i.e. the learner's own manner of controlling multiple acoustic factors, was visualized by *triangular representation*. In assessment experiments, the accordance between the visualized habits and the pronunciation proficiency of the individual learners was analyzed. Results showed us that they were highly correlated with each other, which clearly indicates the validity of the proposed visualization method.

KEYWORDS: English word stress, Pronunciation habit, HMM, English CAI

INTRODUCTION

While an acoustic event called 'word accent' is said to have the same linguistic role between English and Japanese, its acoustic realization differs between them. Japanese word accent is represented by an $\boldsymbol{F_0}$ contour of the word and English one is characterized by four factors of vowel quality, power, $\boldsymbol{F_0}$, and duration. A previous study in phonetics[1] showed that

Japanese learners tend to generate English word accent mainly by manipulating $\boldsymbol{F_0}$. This pronunciation habit should be regarded as inevitable and peculiar to Japanese. And the strength of the habit can be used as an index to estimate English pronunciation proficiency.

In our previous work, a method of detecting a stressed syllable in an English word utterance was proposed[2]. In this method, all the syllables were classified into several dozens of groups in terms of their structural and positional attributes and each syllable group was acoustically modeled by using HMMs (Hidden Markov Models) with duration control. Feature vectors of the HMMs consisted of cepstrum-, power-, and $\boldsymbol{F_0}$-related parameters. It follows that the stress detection was done by using scores of the above four factors. And the estimation and the visualization was realized in this paper by using this detection technique.

ESTIMATION OF THE PRONUNCIATION HABITS AND THEIR VISUALIZATION

Automatic detection of the stressed syllables As shown in Section **1**, the detection of a stressed syllable in an input word was done based upon the maximum likelihood criterion. Here, a syllabic transcription of the word and the number of syllables were treated as *given*. After acoustic feature extraction, syllable boundaries were automatically detected and each syllable was matched with a stressed or unstressed HMM. The stress pattern candidate which produced the highest word-level score was identified as the *correct* stress pattern.

Estimation of the pronunciation habits In the HMM matching procedure, likelihood score f at time t and state i is calculated as

$$f(i,t) = \max_{j,\tau} \left[f(j, t-\tau) a_{ji} d_i(\tau)^{\phi} \prod_{k=1}^{\tau} \prod_{s=1}^{3} b_i^s(y_{t+1-k}^s)^{\rho_s} \right], \quad \left(\sum_{s=1}^{3} \rho_s = 3.0, \quad \rho_s \geq 0.0 \right). \quad (1)$$

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 respectively. y_t^s is a feature sub-vector, which is representing cepstrum-, power-, or $\boldsymbol{F_0}$-related parameters. And ϕ and ρ_s are weighting factors for $d_i(\tau)$ and $b_i^s(y_t^s)$ respectively. This equation can be interpreted as a formula producing the likelihood score $f(i,t)$ by multiplying sub-scores $d_i(\tau)$ and $b_i^s(y_t^s)$ with their weighting factors ϕ and ρ_s. In other words, the score is obtained by integrating the sub-scores of the acoustic observations on spectrum($b_i^1(y_t^1)$), power($b_i^2(y_t^2)$), tone($b_i^3(y_t^3)$), and tempo($d_i(\tau)$) with their adequate weights. It should be noted that the above four sub-scores directly correspond to the four acoustic factors required for the stress/unstress identification.

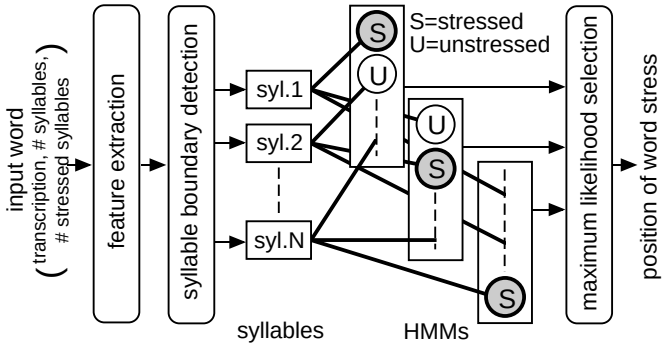

Figure 1: Automatic word stress detection

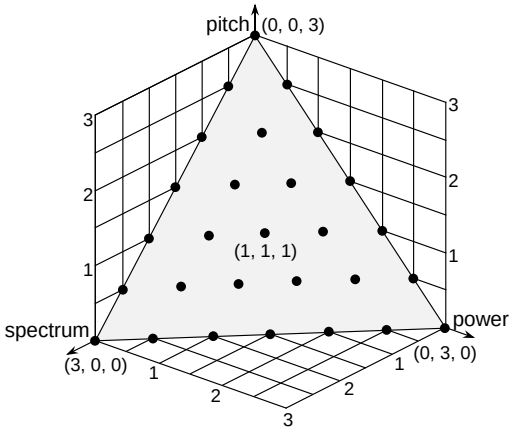

Figure 2: Distribution of the weighing factors

In training the HMMs, all the weighing factors, ϕ and ρ_s, were set to be 1.0. This condition, however, dose not require the factors to be set to 1.0 in detection phase. And it can be easily supposed that the factors $(\rho_1, \rho_2, \rho_3, \phi)$ which give us the highest detection performance is *not* (1.0,1.0,1.0,1.0) especially when the speaker is a non-native speaker. In this paper, we examined modifying the factors only in detection phase. And the optimal combination of the factors were thought to reflect the acoustic features dominantly used for the word stress generation, i.e. the prosodic pronunciation habits of the individual learners.

Visualization of the estimated habits The visualization of the habits can be done by visualizing the relations between the weighting factors and the detection rate[2]. However, the visualization of a five-dimensional function is difficult on a two-dimensional plane. Therefore in this study, only the factors ρ_s were focused upon because they are satisfying a condition $\sum \rho_s = 3.0$ $(\rho_s \geq 0)$ and are found on a plane shown in **Figure 2**. 28 dots in the figure were used for the visualization, where the detection rate at each dot is represented by a shade of colors. Examples are shown in **Figure 3**, where darker colors mean higher rates.

ASSESSMENT OF THE PROPOSED METHODS

Procedures of the assessment experiments Sixty categories of (un)stressed syllables were separately modeled by the HMMs with training data of approximately 3,500 word utterances spoken by one British speaker. For the assessment experiments, English word utterances spoken by seven Japanese learners (**J-1** to **J-7**) with various levels of pronunciation proficiency were prepared as well as word utterances by American speakers (**N-1** to **N-7**).

Firstly by using Japanese samples, the pronunciation proficiency of each learner was rated by English teachers. Secondly, the pronunciation habits were estimated and visualized for each learner. Here, the habits were also calculated for American speakers. Finally, the assessment of the proposed methods was done based upon two comparisons. One was between the visualized habits of Japanese and those of Americans and the other was between the visualized habits of Japanese and their pronunciation proficiency rated by the teachers.

Results of rating the English pronunciation proficiency The English pronunciation proficiency was rated by four English teachers using a five-degree scale (1 to 5). The averaged scores are listed in **Table 1**. According to the table, the learners can be divided into 5 groups; **J-4**→**J-3**→**J-1**→**J-2/5**→**J-6/7** in descending order of the pronunciation proficiency.

Results of visualizing the pronunciation habits and discussions Examples of the visualized habits of four native speakers and four Japanese learners are shown in **Figure 3**. Numbers are the detection rates at the corresponding dots. Double circles are the minimum and maximum of the detection rate. Two sets of three single circles indicate the minimum and maximum of the *averaged* rates over the neighboring three circles.

Firstly, the global differences between the native and Japanese habits are examined. Here, the distances between each vertex (pitch, spectrum, or power) and the center of the three single circles for the averaged maximum are analyzed using ANOVA. Locations of the above centers of Japanese and native speakers are shown in **Figure 4**. Results show us that the distance from the spectrum vertex is significantly shorter in native speakers $(F_{(1,11)} = 6.55, p = 2.65 \times 10^{-2})$ and that the distance from the power vertex is also shorter in native speakers $(F_{(1,11)} = 18.77, p = 1.19 \times 10^{-3})$. However, the distance from the pitch vertex is found to be significantly shorter in Japanese speakers $(F_{(1,11)} = 34.00, p = 1.14 \times 10^{-4})$. These results are very accordant with the findings previously reported in phonetics.

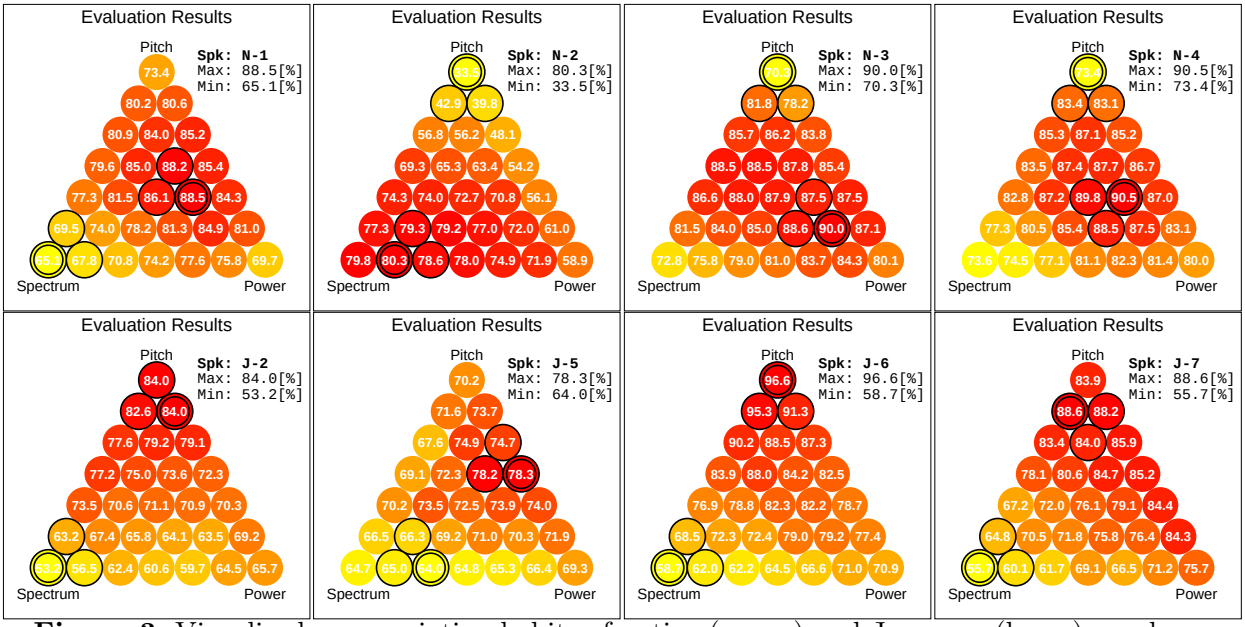

Figure 3: Visualized pronunciation habits of native (upper) and Japanese (lower) speakers

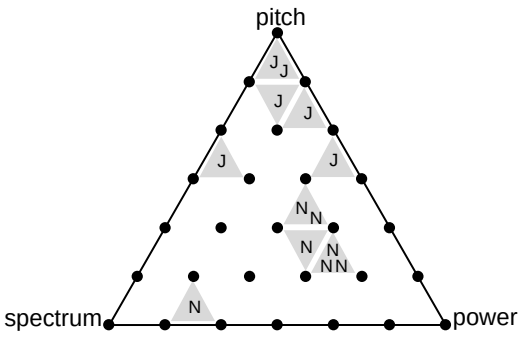

Figure 4: Distribution of the neighboring three circles for the averaged maximum

Table 1: Averaged pronunciation proficiency

J-1	J-2	J-3	J-4	J-5	J-6	J-7
2.90	2.60	3.20	4.45	2.58	1.55	1.55

Secondly, the comparison is done between the habits of Japanese learners and their scores of the pronunciation proficiency. Under the assumption that the degree of being Japanese in pronouncing English words can be estimated by a distance of the position of the three single circles from the pitch vertex, **J-1** to **J-7** will be re-arranged in ascending order of being Japanese as **J-4**→**J-3**→**J-1**→**J-7**→**J-2**→**J-6**. This order is highly correlated with the descending order of their pronunciation proficiency **J-4**→**J-3**→**J-1**→**J-2/5**→**J-6/7** rated by the four English teachers, which is described in the previous section.

CONCLUSIONS

This paper proposed methods to estimate the pronunciation habits and visualize the estimated habits. The estimation was done by using a stressed syllable detector and the visualization was realized by introducing the triangular representation, where acoustic observations are represented abstractly. Assessment experiments showed us that the visualized habits and the pronunciation proficiency of learners were highly correlated. As future works, we are planning to examine how to generate effective instructions based upon the visualized habits and have English teachers assess the proposed methods in a real classroom.

REFERENCES

1. 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).
2. 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).