VISUALIZATION OF PRONUNCIATION HABITS BASED UPON ABSTRACT REPRESENTATION OF ACOUSTIC OBSERVATIONS

Nobuaki MINEMATSU[†] *and Seiichi NAKAGAWA*[‡]

[†] University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN
[‡]Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi, 441-8580 JAPAN
mine@gavo.t.u-tokyo.ac.jp / 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 learners of a language tend to transmit the linguistic information in a manner not of the target language but of 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 a pronunciation habit which is inevitable and peculiar to Japanese learning English. The current study aimed at developing a scheme to evaluate the pronunciation proficiency, where the strength of the pronunciation habit was estimated and used for the evaluation. In this paper, the following three issues of the development were described; 1) the estimation of the pronunciation habits of individual learners by using a stressed syllable detection technique, 2) the visualization of the estimated habits based upon the abstract representation of acoustic observations, and 3) the evaluation of the English pronunciation proficiency by referring to the visualized habits. As a result, the high accordance was observed between the visualized habits and the pronunciation proficiency of individual learners rated by four English teachers. This clearly indicates the validity and effectiveness of our proposed scheme.

1. INTRODUCTION

Recent advances in speech recognition techniques make it possible to develop CALL systems especially for pronunciation learning. To built a CALL system for English pronunciation, the following three evaluating schemes, which are based upon linguistic units of different length, should be prepared; 1) evaluating the adequacy of phonemic features in the learner's utterances, 2) evaluating the adequacy of stress patterns in his/her word utterances, and 3) evaluating the adequacy of intonation over his/her sentence utterances. Since the speech recognition techniques have been basically devised only to identify phonemes, however, the acoustic features irrelevant to the identification are often discarded. It means that only the segmental features are extracted from the signals and the prosodic features are generally ignored. Therefore, the application of the speech recognition techniques without any extension can realize only the first evaluation above. In previous studies, we can find some works which emphasize the importance to learn the pronunciation in terms of prosody. English words with wrong stress patterns are more difficult for native speakers to accept than those with wrong phonemic features so long as the wrong features are found consistently in the utterances[1]. Native speakers perceive the word stress as an essential information source to identify isolated word utterances[1]. The assignment of incorrect stress patterns to words degrades the transmission of segmental information[2]. These works indicate that the pronunciation learning in terms of prosody, which corresponds to the second and the third evaluations, is so important as that in terms of phonemic features.

The immediate use of the speech recognition techniques in building a CALL system causes another problem. As is well-known, in the development of speech recognizers, language-dependent acoustic phenomena tend to be ignored. This is mainly because a speech recognizer with language-dependent modules should become costly to build and difficult to modify into another language's recognizer. On the contrary, it is evident that teaching a language to a learner should carefully focuses on the acoustic phenomena specific to the target language of the learning and to the native language of the learner. Considering the state of the current speech recognition techniques, the immediate application of the techniques can hardly realize a language-dependent teaching scheme, e.g. a scheme designed especially to teach English to Japanese.

As a matter of course, the language-dependent acoustic phenomena can be found in English spoken by Japanese. One example can be seen in word utterances. While an acoustic event called 'word accent' or 'word stress' is said to have almost the same linguistic role between English and Japanese, its acoustic realization differs between the two. The Japanese word accent is represented by an $\boldsymbol{F_0}$ contour of the word[3][4] and the English one is said to be characterized by four factors of vowel quality, power, $\boldsymbol{F_0}$, and duration[5]–[9]. And according to a previous study in phonetics[10], Japanese learners tend to generate the English word accent mainly by manipulating only $\boldsymbol{F_0}$ in the utterances, which is the very manner of the accent generation of the Japanese language. This language-dependent phenomenon can be viewed as a *prosody-based* pronunciation habit which is *inevitable and peculiar to Japanese.* Although the strength of the habit observed in a learner's utterances can surely be a good measure of the pronunciation proficiency of the learner, the current speech recognition techniques can scarcely estimate the habit because it is highly related to prosody and specific to Japanese as well.

In our previous work, a method of detecting a stressed syllable in an English word was proposed by extending the speech recognition techniques[11][12]. In this method, all the kinds of stressed syllables and unstressed ones were grouped into several dozens of classes in terms of the accentual, structural, positional, and/or contextual attributes of the syllable and each syllable class was acoustically modeled as a duration controlled HMM. Using the HMMs, a stressed syllable detector was implemented. By modifying this detector, the present study aimed at developing a scheme to evaluate the pronunciation proficiency, where the above pronunciation habit was estimated and used for the evaluation. In this paper, the following three issues of the development were described[13]; 1) the estimation of the pronunciation habits of individual learners by modifying the stressed syllable detector, 2) the visualization of the estimated habits based upon the abstract representation of acoustic observations, and 3) the evaluation of the English pronunciation proficiency by referring to the visualized habits.

2. DETECTION OF THE STRESSED SYLLABLE IN SPOKEN WORDS

Since the estimation and visualization of the pronunciation habits were implemented by using the stressed syllable detection technique proposed by the authors[11][12], the technique is briefly described here.

2.1. Acoustic Analysis and Feature Parameters

Speech samples were digitized with 12 kHz and 16 bit sampling and the 14-th order LPC analysis was carried out using 21.3 msec frame length and 8.0 msec frame rate. $\boldsymbol{F_0}$ and power were also extracted with the same rate and, after being transformed to the logarithmic scale, they were biased to have zero as mean values over each sample. When modeling the (un)stressed syllables, $\boldsymbol{F_0}$ values for unvoiced segments were required. For these segments, $\boldsymbol{F_0}$ values were estimated by performing the liner interpolation between the preceding voiced segment and the succeeding one and the smoothing of the interpolated $\boldsymbol{F_0}$ contours.

We can find some previous works which attempted to model the stressed syllables and the unstressed ones acoustically[5]−[9]. $\boldsymbol{F_0}$ and power were introduced to represent the (un)stressed syllables[5]−[7]. Duration of the syllable was also considered an important factor to distinguish the stressed syllables from the others[8][9]. Besides the prosodic features above, we can find several works which used the segmental features for the acoustic modeling. In [5] and [7], coarse spectral envelopes were used as one of the acoustic parameters, and in [6], formant-based analysis was performed to estimate the vowel quality. After these works, in the current study, the following three feature streams were used to make a parameter vector; 1) 1 to 4 dimensions of LPC mel cepstrum coefficients and their derivatives, 2) power and its derivative, 3) $\boldsymbol{F_0}$ and its derivative. It should be noted that cepstrum coefficients were calculated after CMN (Cepstrum Mean Normalization) to improve the performance. Using the three streams above, the (un)stressed syllables were acoustically modeled by using the duration controlled HMMs.

2.2. Modeling the Stressed Syllables And the Unstressed Syllables

English is said to have as many as approximately ten thousand different syllables[14]. Therefore in this paper, the English syllables were grouped into syllable *classes* in terms of their accentual, positional, and/or structural attributes to be modeled by the HMM.

- grouped into 2 classes; the stressed syllables and the unstressed ones. The *accentual* attribute of the syllable is only considered.
- grouped into 6 classes; $\mathrm{S_H}$, $\mathrm{S_T}$, and $\mathrm{S_O}$ separately for the stressed syllables and the unstressed ones, where $\mathrm{S_H}$ and $\mathrm{S_T}$ denote a syllable at the head and one at the tail of a word respectively, and $\mathrm{S_O}$ indicates a syllable at the other parts of the word. In this case, the *accentual* and *positional* attributes of the syllable in the word are introduced into the HMMs.
- grouped into 16 classes; V_S, $\mathrm{C}V_\mathrm{S}$, $V_\mathrm{S}\mathrm{C}$, $\mathrm{C}V_\mathrm{S}\mathrm{C}$, V_L, $\mathrm{C}V_\mathrm{L}$, $V_\mathrm{L}\mathrm{C}$, and $\mathrm{C}V_\mathrm{L}\mathrm{C}$ separately for the stressed syllable and the unstressed ones, where V_S and V_L represent a short vowel and a long vowel respectively and C means a sequence of consonants. In this grouping, the *accentual* and *structural* attributes of the syllable are integrated into the HMMs.
- grouped into 48 classes; the above three attributes are considered. It follows that a syllable model is built for each combination of its accentuation, its position in the word, and its syllabic structure.

The use of the *positional* attribute in the second and the forth groupings is derived from the following preliminary discussion. An observed $\boldsymbol{F_0}$ contour generally shows a rising pattern at the beginning of an utterance and a falling pattern at the end[3], which is language independent[4]. And this is the case even when the utterance is an isolated word[4]. It means that the first and the last syllables in the word utterance should be separately modeled at least in terms of its $\boldsymbol{F_0}$ contour.

All the experiments of this paper were carried out by using the duration controlled CDHMMs with six states and four distributions, which were built by the forth grouping, i.e. 48 classes. It should be noted that a PDF was comprised of a single Gaussian distribution with a full covariance matrix and that the correlation between any two of the three feature streams, cepstrum-, power-, and pitch-related parameters, was assumed to be zero in the covariance matrix.

Throughout the experiments of this paper, as listed in **Table 1**, polysyllabic words of ATR British English

Table 1: English word samples for training the HMMs

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

Table 2: English word samples for testing the proposed method

set	#spk	native lang.	vocab. size	#words
A	7	**A**merican	381	546
J	7	**J**apanese	60	341

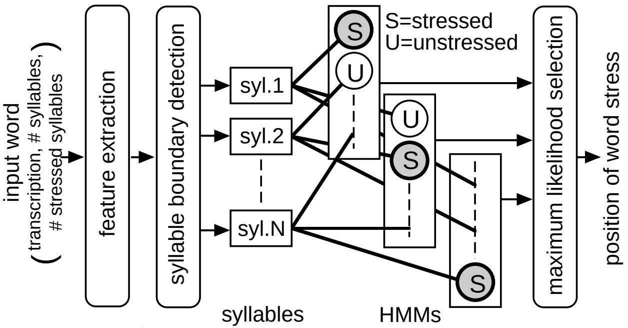
input word
(transcription, # syllables, # stressed syllables)
feature extraction
syllable boundary detection
syl.1
syl.2
syl.N
S
U
S=stressed
U=unstressed
maximum likelihood selection
position of word stress
syllables
HMMs

Figure 1: Automatic detection of the stressed syllable

word database were used as training samples. As for testing samples, speech material of **Table 2** was used. **A** is a set of American English polysyllabic words from Resource Management isolated word database and **J** is a set of English words spoken by Japanese, which were recorded in a soundproof room of our laboratory.

2.3. Automatic Detection of the Stressed Syllable in Word Utterances

The detection of the stressed syllable in an input word was carried out based on the maximum likelihood criterion using a word-level score. As shown in **Figure 1**, the input word was firstly divided into syllables, and then each segmented syllable was matched with its corresponding stressed or unstressed HMM in the candidate stress patterns. The summation of the syllable-level scores made the word-level score. And the position of the stressed syllable HMM in the candidate pattern which produced the highest word-level score was identified as stressed. In the figure, the syllable segmentation was performed based upon the forced Viterbi alignment with English phoneme HMMs. If the input word was spoken by a Japanese learner, the phoneme HMMs adapted to the English spoken by Japanese were used for the alignment. It should be noted that the syllabic transcription of the word, the number of syllables and that of stressed syllables (one throughout the experiments) of the word were all treated as given. Hence, the number of the candidate stress patterns was N for N-syllable input words.

3. ESTIMATION OF THE PRONUNCIATION HABITS

3.1. Reviews on the Viterbi Decoding

In the HMM matching procedure, the likelihood score is approximately obtained using the Viterbi decoding and the score is called the Viterbi score. And at time t and state i, the Viterbi score $f(i,t)$ 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 from state j to i, a duration probability of staying at state i for a period of τ, and an output probability density that vector y_t is generated from state i. And ϕ is a weight for $d_i(\tau)$. As mentioned before, input vector y_t is composed of three streams; cepstrum-, power-, and $\boldsymbol{F_0}$-related parameters. And assuming the correlation between any two of the above three streams to be zero, $b_i(y_t)$ can be written as

$$b_i(y_t) = \prod_{s=1}^{3} b_i^s(y_t^s)^{\rho_s}, \text{ where } \sum_{s=1}^{3} \rho_s = 3.0 \ \ (\rho_s \geq 0.0). \quad (2)$$

y_t^s represents a sub-vector corresponding to one of the above three streams and ρ_s is a weight for $b_i^s(y_t^s)$. Finally, we can get the Viterbi score $f(i,t)$ again as

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

This equation can be interpreted as a formula producing the Viterbi score $f(i,t)$ by multiplying sub-scores $d_i(\tau)$ and $b_i^s(y_t^s)$ with their individual weights ϕ and ρ_s. It means that the score is obtained by integrating the sub-scores of the acoustic observations of 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. These four sub-scores directly correspond to the four factors needed for characterizing the (un)stressed syllables, which was described in Section **1.**.

In modeling the syllable classes, all the weights, ϕ and ρ_s, were set to be 1.0. However, it can be supposed that the combination of weights $(\rho_1, \rho_2, \rho_3, \phi)$ which gives us the highest detection rate should *not* be (1.0,1.0,1.0,1.0) though the (un)stressed syllable HMMs were trained with all the weights being 1.0. This is because of acoustic mismatches between training samples and testing ones and, due to the *inevitably* observed acoustic distortions, it should be especially the case when the testing samples are spoken by non-native speakers. And the adequate modification of the stream weights should improve the detection performance. Here, the increase of a weight of a specific acoustic feature is mainly interpreted to emphasize the likelihood score of the acoustic feature in the detecting procedures. In other words, modifying the stream weights can be considered to be adapting the hearing characteristics of a computer's ears. And the optimal combination, which maximizes the detection rate, is thought to reflect the acoustic features dominantly used for the word accent generation, i.e. the pronunciation habits of individual learners[12]. **Figure 2** shows the above interpretation of the modification of the feature stream weights. In this figure, the authors' expectations are also drawn. While the increase of the

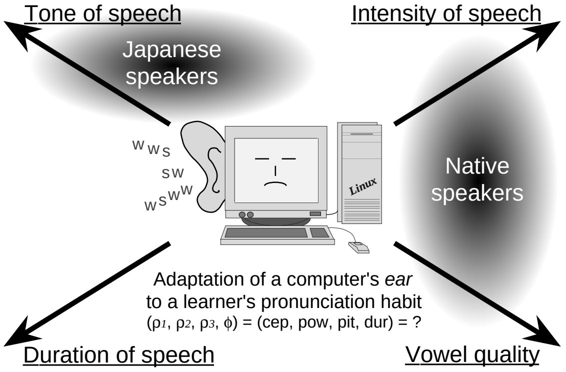

Figure 2: Adaptation of a computer's ear to a learner's own pronunciation habit

pitch weight is expected to improve the detection rate with English words spoken Japanese, that of the power weight and the vowel quality weight is thought to do with words spoken by native speakers.

3.2. Estimation of the Pronunciation Habits

The estimation of the optimal weight was intentionally carried out by using a *greedy* strategy. The reason for that will be described in Section **4.**. Here, out of a pre-defined set of combinations, the combination giving the highest performance was obtained by calculating the detection rate for each combination. The pre-defined set is as follows.

- A weight for duration, ϕ, was changed from 0.0 to 20.0 with a step of 0.5. Then, ϕ had 41 varieties.
- For each value of ϕ, weights for the other parameters, ρ_s, were set to be one of the following 28 combinations, where $\sum \rho_s = 3.0$ and $\rho_s \geq 0$.

$$(\rho_1, \rho_2, \rho_3) = \begin{cases} (1.0, 1.0, 1.0), \\ (1.5, 0.5, 1.0), (1.5, 1.0, 0.5), (0.5, 1.5, 1.0), \\ (1.0, 1.5, 0.5), (0.5, 1.0, 1.5), (1.0, 0.5, 1.5), \\ (0.0, 1.5, 1.5), (1.5, 0.0, 1.5), (1.5, 1.5, 0.0), \\ (2.0, 0.0, 1.0), (2.0, 1.0, 0.0), (0.0, 2.0, 1.0), \\ (1.0, 2.0, 0.0), (0.0, 1.0, 2.0), (1.0, 0.0, 2.0), \\ (2.0, 0.5, 0.5), (0.5, 2.0, 0.5), (0.5, 0.5, 2.0), \\ (2.5, 0.0, 0.5), (2.5, 0.5, 0.0), (0.0, 2.5, 0.5), \\ (0.5, 2.5, 0.0), (0.0, 0.5, 2.5), (0.5, 0.0, 2.5), \\ (3.0, 0.0, 0.0), (0.0, 3.0, 0.0), (0.0, 0.0, 3.0). \end{cases} \quad (4)$$

Consequently, an input word was matched with a concatenation of the (un)stressed syllable HMMs with 41×28=1148 varieties of the stream weight combinations. And the combination giving the highest detection rate for several dozens of word utterances should be obtained as the pronunciation habit of the learner.

4. VISUALIZATION OF THE ESTIMATED HABITS

4.1. Preliminary Discussions on Visualization

In the previous section, the pronunciation habit was defined as the optimal weight combination in the weight space. However, the detection rates with non-optimal weight combinations can also be of great help to provide a learner with instructions on his/her pronunciation habit. For example, the following issues can be dealt with only by considering the entire distribution of the detection rate in the weight space; 1) the weight combinations showing the comparable performance to that of the optimal point can be given to learners, 2) the point corresponding to the lowest performance can also be considered to be one of the aspects of the learner's pronunciation habit, and 3) when using the optimal point only, it is impossible to examine how valid the location of the learner's optimal point is in the native speakers' weight space. Those are why the authors did a greedy method, where the weight space were quantized adequately and the detection rate of each point or centroid was calculated greedily. In the following section, a method of visualizing the entire distribution of the detection rate is devised.

4.2. Visualization of the Estimated Habits

The complete visualization of the estimated habit requires a method of representing the entire distribution of the detection rate in the space of four weights, ϕ and $\{\rho_s\}$. And this task is surely very difficult to do on a two-dimensional plane. In this section, for the following two reasons, the visualization of the relation between the detection rates and $\{\rho_s\}$, i.e. weights for cepstrum, power, and $\boldsymbol{F_0}$, is focused upon.

- Weighting operations are done differently between duration and the other three parameters, which are shown in Equation (**3**).
- Unlike cepstrum, power, and $\boldsymbol{F_0}$ parameters, duration for a state is used in the HMM without any normalization.

As shown in Equation (**2**), the above three weights are satisfying a condition $\sum \rho_s = 3.0$ ($\rho_s \geq 0$). Therefore, 28 combinations in Equation (**4**) can be plotted on one plane, which is shown in **Figure 3**. And by representing the detection rate of each dot (weight) using different colors, the visualization of the relation between the detection rates and $\{\rho_s\}$ under a specific value of ϕ can be realized. This representation will be called *triangular representation* in the rest of the paper. Although the triangular representation enables learners to know the pronunciation habit visually, it provides them with a triangle per duration weight, namely, 41 triangles all together. Then, a method of integrating the 41 triangles into one representative pattern is required. Since the triangular representation currently deals with the other three parameters than duration, the integration of the 41 triangles can be done by calculating an expectation pattern of the triangles along an axis of duration. And this integration is shown in **Figure 4**. Firstly, the averaged detection rate on a triangle of $\phi = \Phi$ is calculated as $d(\Phi)$. Secondly, a weighting factor assigned to the $\phi = \Phi$ triangle for the expectation operation is defined as $w(\Phi) = d(\Phi)/\sum_{\phi} d(\phi)$. Finally using these weights, the representative pattern of the 41 triangles is produced by the expectation operation along an axis of duration. Here, detection rate $S(\rho_1, \rho_2, \rho_3)$ on the

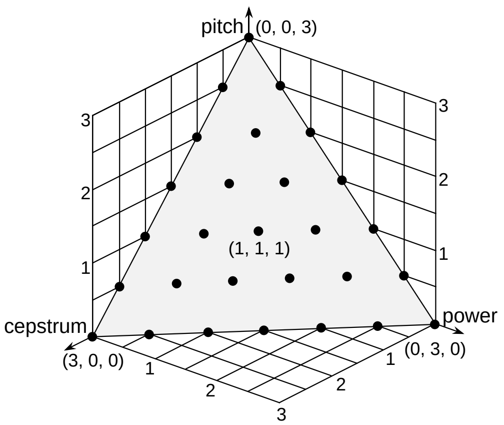

Figure 3: Distribution of the weight combination of $\{\rho_s\}$

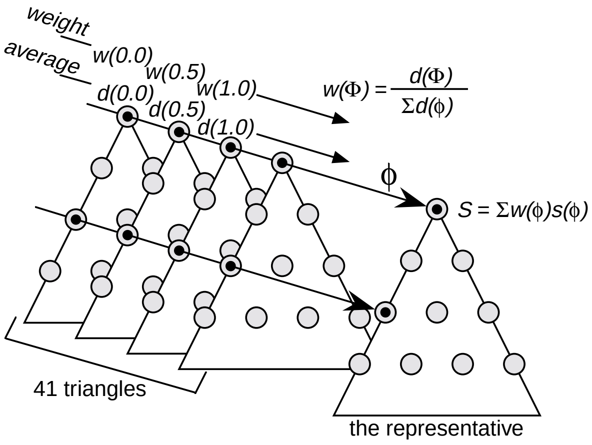

Figure 4: Integration of the 41 triangles into the representative pattern

representative triangle is calculated as

$$S(\rho_1, \rho_2, \rho_3) = \sum_{\phi} w(\phi) s(\rho_1, \rho_2, \rho_3, \phi), \qquad (5)$$

where $s(\rho_1, \rho_2, \rho_3, \phi)$ means the detection rate at weight $(\rho_1, \rho_2, \rho_3, \phi)$. **Figure 5** shows an example of the representative pattern. Here, a shade of colors of the circles corresponds to the height of the detection rate. Darker and lighter circles indicate higher and lower rates respectively. Numbers in the circles are the expected detection rates. Two double circles indicate the maximum and the minimum of the expected detection rates, henceforth the maximum/minimum circle. Two sets of three single circles mean the maximum and the minimum of the average of the expected detection rates over the *neighboring* three circles, henceforth the maximum/minimum neighboring circles.

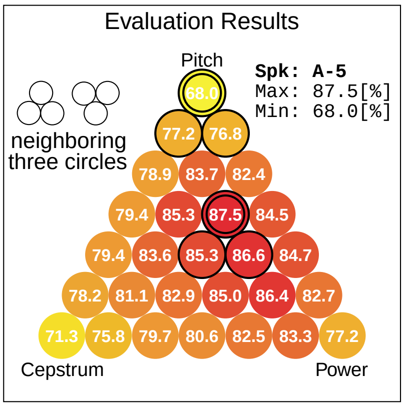

Figure 5: An example of the representative triangle

The triangular representation can show a learner the (un)balance of his/her controlling vowel quality, power, and $\boldsymbol{F_0}$, and that without any presenting acoustic observations immediately. In some CALL softwares, the acoustic observations such as speech waveforms, $\boldsymbol{F_0}$ curves, power curves, and spectrum patterns are separately provided with learners. However, most of the learners are easily supposed to be unfamiliar with the above immediate representation of the acoustic observations. Therefore, the immediate visualization may have only small effects on pronunciation learning. Even if the learners have enough knowledge on the observations, the immediate visualization must be of little use when the acoustic features are *separately* presented to the learners. This is because the word stress of English is generated by adequately controlling the multiple acoustic factors *simultaneously*, in other words, adequately balancing the multiple factors. As mentioned above, the triangular representation shows the learner's manner of the control based upon the *abstract* representation of the acoustic observations and the *integrated* representation of the multiple factors. Hence, the effectiveness of using this representation is expected to be extremely high.

5. ASSESSMENT OF THE PROPOSED METHOD

5.1. Procedures of the Assessment Experiments

English word samples spoken by seven Japanese learners with various levels of English pronunciation proficiency were prepared. They are listed as **J** in **Table 2**. By using these samples, the pronunciation proficiency of the individual learners was firstly rated by four English teachers. After that, the representative triangle was automatically generated separately for each learner. And the representative triangle was also made for each native speaker of **A**. The assessment of the proposed method was conducted based upon two comparisons. One is the comparison between the representative triangles of the Japanese learners and those of the native speakers and the other is that between the triangles of the Japanese learners and the pronunciation proficiency of the learners rated by the English teachers. In other words, the first comparison examines whether the proposed method can detect the differences of the accent generation manner between Japanese learners and native speakers. And the second one investigates whether the method can detect the differences among Japanese learners.

5.2. Rating of the English Pronunciation Proficiency of Japanese Learners

The rating of the English pronunciation proficiency was done by four English teachers using a five-degree

Table 3: Averaged pronunciation proficiency rated by four English teachers

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

scale (1 to 5). The averaged scores over the teachers are shown in **Table 3** separately for each learner. It should be noted that the scores in the table contain the proficiency to speak English phonemes correctly as well as that to generate the English word stress adequately. According to the table, the learners can be divided into five groups;

J-4→J-3→J-1→J-2/5→J-6/7

in descending order of the pronunciation proficiency.

5.3. Results of the Pronunciation Habit Estimation and Discussions

The representative triangles of seven native speakers (**A**) and seven Japanese learners (**J**) are shown in **Figure 7**. When the highest score in a representative triangle is absolutely low, it indicates that the adaptation based upon the weight modification does not work well enough for the speaker. Therefore in the following discussions, only the speakers whose highest detection rate is over 80 % are considered, which means that the triangle of **J-5** is ignored below.

Firstly, the differences between the native triangles and the Japanese ones are examined. The locations of the maximum neighboring circles of Japanese learners and native speakers are shown in **Figure 6**. Here, the distances between each vertex (pitch, cepstrum, or power) and the center of the maximum neighboring circles were analyzed using ANOVA. Results showed that the distance from the spectrum vertex was significantly shorter in native speakers ($F_{(1,11)} = 6.55, p = 2.65 \times 10^{-2}$) and that the distance from the power vertex was also significantly shorter in native speakers ($F_{(1,11)} = 18.77, p = 1.19 \times 10^{-3}$). However, the distance from the pitch vertex was found to be significantly shorter in Japanese learners ($F_{(1,11)} = 34.00, p = 1.14 \times 10^{-4}$). These results clearly indicate that the Japanese learners tend to generate the English word accent by means of the accent generation manner of the Japanese language. As told in Section **1.**, this tendency is also reported in a previous study in phonetics[10].

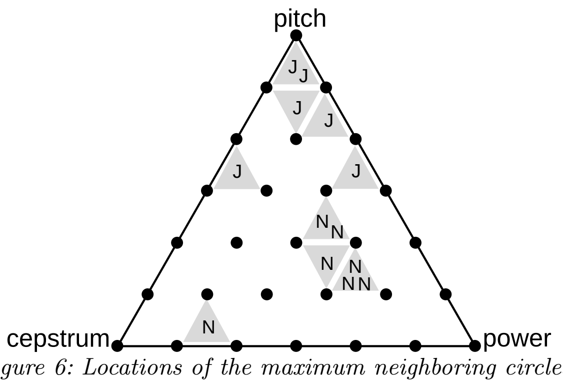

Figure 6: Locations of the maximum neighboring circles

We can say that the proposed method is able to detect the differences of the accent generation manner between native speakers and Japanese learners. Another finding, which is related to the minimum neighboring circles, is also obtained from **Figure 7**. The lowest rate in every Japanese learner is found at the cepstrum vertex. This should be due to the following two reasons. One is the incorrect syllabic structure caused by inserting an additional vowel after each consonant, which is very peculiar to the English spoken by Japanese. The other is inadequate pronunciation of phonemes, such as /a/ for /æ/. Considering that the syllable HMMs were built using coarse segmental features, however, the main reason should be the first one. As for native speakers, it is clearly seen that the lowest rates tend to be found near the pitch vertex.

Secondly, the comparison is done between the representative triangles of Japanese learners and their pronunciation proficiency rated by English teachers. Assuming that the degree of being Japanese in speaking English can be estimated by the distance from the pitch vertex to the maximum neighboring circles, the Japanese learners can be arranged in ascending order of the degree of being Japanese in speaking English as

J-4→J-3→J-1→J-7→J-2→J-6.

This order is highly accordant with the descending order of their English pronunciation proficiency rated by four English teachers in the previous section;

J-4→J-3→J-1→J-2→J-7/6.

Only **J-2** and **J-7** are arranged reversely between the two arrangements. If the maximum circles (double circles), not the maximum neighboring circles, are considered, however, these two learners should be placed at the same rank. And the detection rate at the maximum circle, namely, with a very large pitch weight, is higher in **J-7**. Considering these results, it may be better to judge **J-7** to be more Japanese than **J-2** in speaking English. To define a *unified and decisive* manner of viewing the representative triangles, the analysis of the triangles using a larger number of speech samples should be required.

5.4. Subjective Assessment of the Proposed Visualization Method

Subjective assessment of the proposed method was done by using questionnaires to seven Japanese learners and five English teachers. Here, they were asked to give us free opinions of the visualized habits. And the learners were the same as those listed as **J** in **Table 2** and one of the five teachers had participated in the rating experiment in Section **5.2.**.

5.4.1. Opinions from Japanese learners

Typical opinions from the learners are shown below.

- The proposed method enables learners to check the (un)balance of controlling the multiple acoustic factors and makes it much easier for the learners to

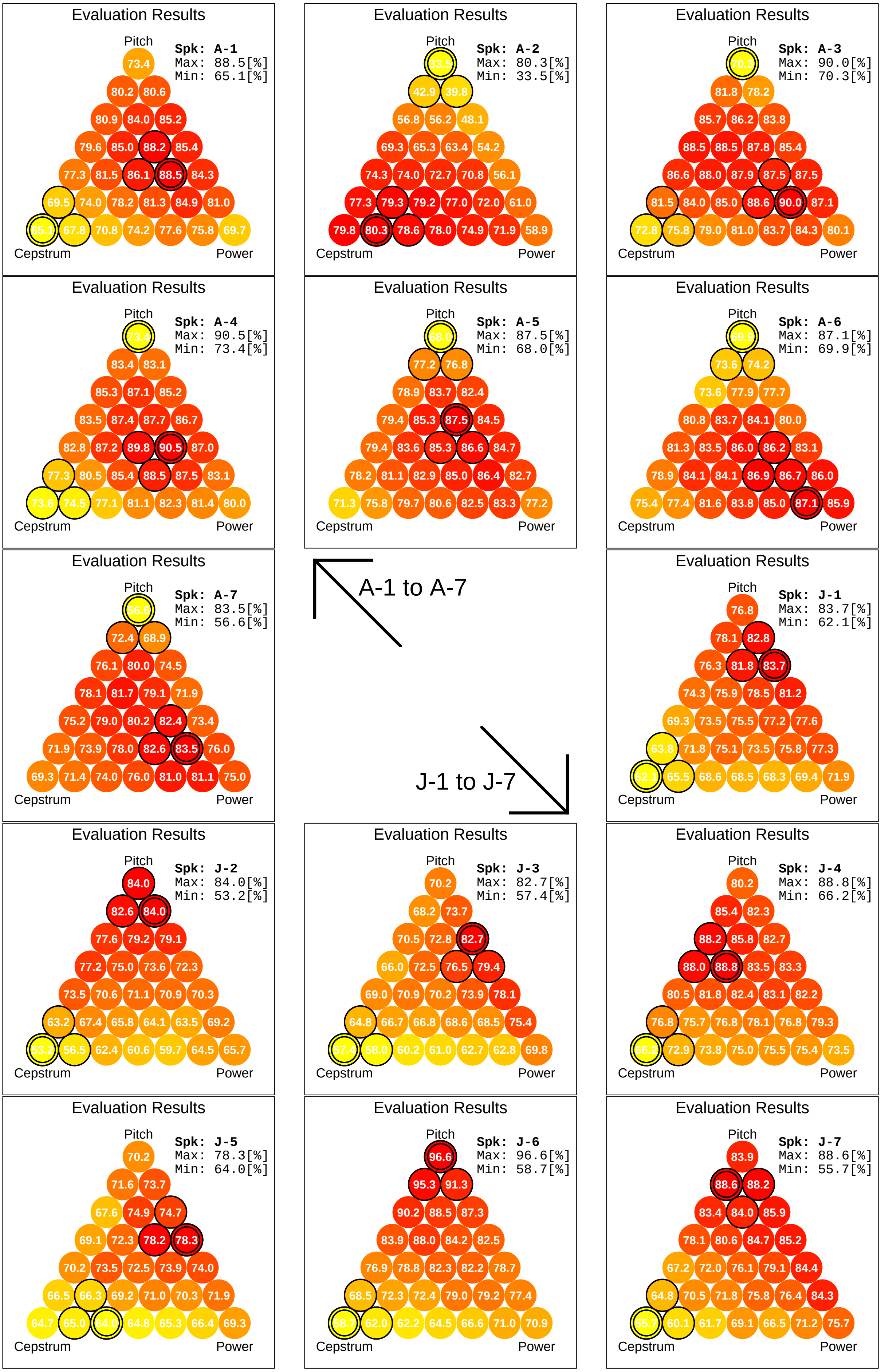

Figure 7: Visualized pronunciation habits of native speakers, **A-1** *to* **A-7***, and Japanese speakers,* **J-1** *to* **J-7**

comprehend their state in the process of pronunciation learning.

- It is easy to compare the manner of learners with that of native speakers. And the proposed method also facilitates the comparison between how the learners are and how they used to be.
- Abstract representation of acoustic features are quite easy to understand. Immediate presentation of waveforms, spectrum patterns, pitch patterns, and so forth are often confusing to learners.

It should be noted that all the Japanese learners who answered the questionnaire agreed that the proposed visualization method is very preferable to the conventional methods of the immediate and separate visualization of acoustic observations.

5.4.2. Opinions from English teachers

Typical opinions from the teachers are listed below.

- It is necessary to integrate the proposed method with a evaluation scheme using phonemic features.
- Evaluation in terms of sentence-level features should be introduced to the proposed method.
- Four factors of vowel quality, power, pitch, and duration are thought to be derived from a good and deep consideration of the acoustic differences between English word accent and Japanese one.
- Triangular representation is quite easy to understand. It has to be confessed that all the English teachers are not familiar with phonetics. So, this kind of abstract representation is quite helpful for teachers as well as for students.
- Quite interesting. But instructions on what to do next should be more concrete.
- English teachers whose mother tongues are not English may want this tool to check and maintain their pronunciation proficiency.
- Triangular representation explicitly with duration information is desired.

The questionnaires inspired the authors to integrate the proposed method with other two evaluation schemes, phoneme-based and sentence-based. As the learners preferred the proposed method above, all the English teachers also gave us favorable comments. The authors consider that these favorable comments ensure the validity of the proposed visualization method.

6. CONCLUSIONS

A method to evaluate Japanese manners of generating English word accent was proposed by using a stressed syllable detector. Here, the learner's manner of the word accent generation was estimated by searching for the optimal combination of weighting factors of the four acoustic features. And the easy-to-understand visualization of the manner was also realized based on the abstract representation of acoustic observations. Assessment experiments showed that the proposed method can evaluate the pronunciation proficiency in accordance with English teachers' evaluation. As future works, we are planning to 1) built the syllable HMMs using a larger number of word samples, 2) devise another visualization method which explicitly shows the learner's manner of controlling the syllable duration, 3) study how to view and analyze the graphically estimated habits to give the final evaluation, 4) integrate the proposed method with other evaluation schemes in terms of phonemic features and sentence intonation, and 5) develop an effective and interactive feedback method to instruct the learner to correct his/her pronunciation habit.

REFERENCES

1. G. Kawai and A. Ishida, “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).
2. A. Cutler and D. Norris, “The role of strong syllables in segmentation for lexical access,” J. Experimental Psychology: Human Perception and Performance, vol.14, pp.113–121 (1988).
3. H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Jpn.(E), vol.5, no.4, pp.233–242 (1984).
4. H. Fujisaki, K. Hirose, and M. Sugito, “Comparison of acoustic features of word accent in English and Japanese,” J. Acoust. Soc. Jpn.(E), vol.7, no.1, pp.57–63 (1986).
5. G. J. Freij, F. Fallside, C. Hoequist, and F. Nolan, “Lexical stress estimation and phonological knowledge,” Computer Speech and Language, vol.4, pp.1–15 (1990).
6. S. Hiller, E. Rooney, J. Laver, and M. Jack, “SPELL: An automated system for computer-aided pronunciation teaching,” Speech Communication, vol.13, pp.463–473 (1993).
7. H. Hamada, S. Miki, and R. Nakatsu, “Automatic evaluation of English pronunciation based on speech recognition technique,” IEICE Trans. vol.E76-D, no.3, pp.352–359 (1993).
8. P. Dumouchel and M. Lening, “Using stress information in large vocabulary speech recognition,” Proc. Montreal Symposium on Speech Recognition, McGill Univ., Montreal, pp.73–74 (1986).
9. P. Lieberman, “Some acoustic correlates of word stress in American English,” J. Acoust. Soc. Am., vol.32, no.4 (1960).
10. 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).
11. N. Minematsu, N. Ohashi, and S. Nakagawa, “Automatic detection of accent in English words spoken by Japanese students,” Proc. European Conf. Speech Communication and Technology, pp.701–705 (1997).
12. Y. Fujisawa, N. Minematsu, and S. Nakagawa, “Evaluation of Japanese manners of generating word accent of English based on a stressed syllable detection technique,” Proc. Int. Conf. Spoken Language Processing, pp.3103–3106 (1998).
13. N. Minematsu, Y. Fujisawa and S. Nakagawa, “Prosodic evaluation of English words spoken by Japanese based upon estimating their pronunciation habits,” Proc. Int. Conf. Speech Processing, pp.439–444 (1999).
14. L. Rabiner and B. H. Juang, “Fundamentals of speech recognition,” Prentice-Hall, New Jersey, pp.436 (1993).