AUTOMATIC DETECTION OF ACCENT IN ENGLISH WORDS SPOKEN BY JAPANESE STUDENTS

Nobuaki MINEMATSU *Nariaki OHASHI* *Seiichi NAKAGAWA*
mine@tutics.tut.ac.jp *ohashi@slp.tutics.tut.ac.jp* *nakagawa@tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Tech.,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441 JAPAN
Tel: +81-532-44-6767, FAX: +81-532-44-6757

ABSTRACT

Acoustic realization of word accent differs among languages. While, in Japanese, it is fully represented by an F_0 contour of a word, English word accent is characterized by power, duration, F_0, vowel quality and so forth. In addition to the difference in syllable structure between the two languages, that in word accent makes it even more difficult for Japanese students to master correct pronunciation of English words. It indicates that the development of an automatic evaluation method of English word accent, as one of English teaching tools, will be helpful especially to Japanese students. In this paper, as the first step to the development, a detection method of accent in English words spoken by Japanese is proposed, where syllable-size HMMs are built using *positional* information of the syllables and adequately detected syllable boundaries are used for the detection. Results of accent detection experiments show 90 % and 93 % as detection rates of Japanese students and native speakers respectively.

1. INTRODUCTION

Evaluation of English words pronounced by foreign students should be based upon the following two indexes[1];

- whether adequate phonemic characteristics are observed in the words, and
- whether the words are uttered with adequate stress patterns and with proper strength of the stress.

And the former and the latter are considered to correspond to the evaluation of segmental features and that of prosodic features observed in the words respectively. It is interesting that English words with wrong stress patterns are reported to be more difficult to be accepted to native speakers than those with wrong phonemic features so long as the wrong features are generated *consistently* in the utterances[2][3].

Further, acoustic realization of English word accent is greatly different from that of Japanese word accent. While, in Japanese, word accent is completely described by an F_0 contour of the word, English word accent is said to be characterized by power, duration, F_0, vowel quality and so forth. Two terms – *accented* and *stressed* – often indicate the same acoustic event in English, but not in Japanese. A previous study reported that Japanese students tend to realize English word accent mainly by increasing F_0 in the word, which is a Japanese manner of accent generation[4]. In addition to the difference in syllable structure between the two languages (almost all the Japanese syllables consist of CV or V), that in word accent is also considered to make it even more difficult for Japanese students to master correct pronunciation of English words.

Although Japanese students are often told to pay more attention to the intonation on their own utterances, the difference in word accent realization between the two languages and the above mentioned perceptual characteristics of native speakers indicate that it is also important (maybe more important) to provide for Japanese students an evaluation scheme of English word accent.

In order to develop an automatic evaluation scheme of stress patterns in English words, the following two steps should be followed;

i) correct detection of a stressed syllable in the word, and

ii) adequate scoring of strength of the stress.

A main objective of this paper is to develop the detection method (step **i**) and to examine its performance using words spoken by native speakers and Japanese students experimentally[5]. Through several modifications of acoustic models built for stressed/unstressed syllables, it is discussed what kind of modification enables the models to capture what kind of acoustic event found in the words, which is considered important especially when applying the detection method to foreign language learning.

In the following section, as preliminary discussion, word accent of English and Japanese is described in terms of its acoustic parameterization. Here, the structural difference of a syllable between the two languages is also referred to. Listening tests conducted to two English teachers for preparing speech samples is described in Section **3**. And in Section **4**, several modeling schemes of (un)stressed syllables are introduced, and they are examined in automatic detection experiments of word accent in Section **5**. Finally in the last section, this paper is briefly summarized.

2. WORD ACCENT OF ENGLISH AND JAPANESE

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 [6], an English syllable has a central vowel and sequences of consonants placed 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 number of kinds of syllable. While there are only one hundred or so kinds in Japanese, English is estimated to have more than two thousand variations.

In previous works on evaluation of English word utterances or on identification of their stress patterns, several kinds of acoustic features and their combinations were used to characterize the word accent. In these studies, coarse spectral envelopes were used[7][8] or formant-based analysis was conducted[1] to estimate vowel quality broadly. *F0* and power were also introduced to measure strength of stressed syllables[8], and furthermore, duration was also considered important to distinguish stressed syllables from others[9]. In the previous studies, however, few works were found where all of these features were effectively integrated. In this study, all of the above parameters are utilized to model (un)stressed syllables. Namely, a parameter vector is composed of 12 elements, 4-th order LPC mel cepstrum coefficients and their derivatives, *F0* and its derivative and power and its derivative. Using this parameterization of speech signals, continuous density HMMs with duration control are adopted for acoustic modeling. In these HMMs, all of the above acoustic features are supposed to be contained.

If the number of kinds of syllable in English were limited as in Japanese, the (un)stress models could be built for individual syllables. As told above, however, a large variation in English syllables requires us to make the models separately for each syllable *class*. And the class-based modeling leads us to use only a small number of dimensions of segmental features, which is four in this case. Several schemes for the syllable clustering are examined taking *structural* or *positional* information of the syllable into account, which will be described in Section **4**.

The above discussions are based upon the premise that an acoustic model for the (un)stress should be built by a unit of syllable. However, it is not impossible to built phoneme-size (un)stress models. If *F0*, power or duration has phoneme dependent distribution to some extent, the phoneme specific models are expected to improve the detection performance. In this paper, in addition to syllable-size models, phoneme-size models are built tentatively and they are compared to each other in Section **5**.

3. LISTENING TESTS BY ENGLISH TEACHERS

To construct the acoustic models for (un)stressed syllables, speech samples with stress labels must be prepared. So, all of the isolated words in Resource Management Database (RM1) which contained more than one syllable were presented to two English teachers through headphones, one of whom was a native speaker and the other was Japanese. In this listening test, they were asked to locate stressed syllables in the words irrespective of their lexical knowledge. This was because some of the words were uttered with illegal word accent though all the speakers of the database were native. Additionally, they were requested to evaluate strength of the stress using a scale of five degrees. After the listening test, speech samples satisfying the following conditions were extracted;

A) the number of stressed syllables in the word is one,

B) the two teachers indicate the same syllable as the stress location of the word, and

C) the evaluation scores of the stress is larger than or equal to a given threshold.

Out of the extracted samples, 708 words spoken by 107 male speakers were used for training, to all of which, using a phoneme-based speech recognizer developed in our laboratory with TIMIT database, forced Viterbi alignment was carried out to detect syllable boundaries automatically. And other 705 words spoken by 7 male speakers were extracted for testing native speakers' utterances, which satisfied only the conditions **A**) and **B**). Approximately 80 % of more-than-one-syllable words in the RM1 database were used for training and testing.

Another listening test was carried out to the same teachers using English words spoken by Japanese students. For this test, 60 English fundamental words containing more than one syllable were prepared, some of which were included in the RM1 database and the others not. And they were uttered by 7 students to be digitized with a DAT recorder. After the listening test, 351 words were obtained for testing Japanese students' utterances, which also satisfied only the conditions **A**) and **B**). Through all the training and the testing procedures below, the syllable which was indicated as *stressed* by both the teachers is treated as correct position of word accent. As told above, it was sometimes different from the lexically correct position of accent of the word.

Since the evaluation scores are naturally supposed to have a bias caused by the subjective difference in the scoring principle between the teachers, they are used in discussion of experimental results after proper normalization.

4. MODELING STRESSED/UNSTRESSED SYLLABLES

As mentioned in Section **2**, all the syllables observed in the training data have to be clustered into some classes for acoustic modeling. In this study, the following clustering schemes are experimentally investigated.

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

b) clustered into 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 and V_L denote a short and a long vowel respectively and C denotes a sequence of consonants the length of which is more than or equal to one. In this clustering, *structural* information of a syllable is integrated into the HMMs.

c) clustered into 6 classes; S_H, S_T, and S_O separately for stressed and unstressed syllables, where S_H and S_T denote a syllable at the head and at the tail of a word respectively, and S_O indicates a syllable at the other parts of the word. In this case, *positional* information of a syllable in the word is introduced into the HMMs.

The third clustering scheme is derived from the following consideration. An observed F_0 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[10]. It means that the first and the last syllables in a word should be separately modeled at least in terms of its F_0 contour. Furthermore, as told in Section **2**, phoneme-size stress models are examined tentatively as shown below.

d) clustered into 42 classes; 12 HMMs and 6 HMMs are built for stressed and unstressed vowels respectively. As for consonant models, a single model is made for each consonant irrespective of its accentuation, which produces 24 HMMs. It should be noted that the phoneme-size HMMs use the same parameter vectors as the syllable-size ones do, namely, not with higher dimensions of cepstrum coefficients. This is because word accent is considered to be poorly represented by fine segmental features compared to that by prosodic features and coarse segmental features.

Through all the experiments, speech materials are digitized with 12 kHz and 16 bit sampling and acoustic analysis is performed using 21.3 msec frame length and 8.0 msec frame rate. F_0 and power are also extracted with the same rate and, after being transformed to logarithmic scale, they are normalized to have zero as averaged values over each utterance. When building the models, F_0 values for unvoiced segments are required. For these segments, F_0 values are estimated by liner interpolation of the preceding/succeeding voiced segments.

5. AUTOMATIC DETECTION EXPERIMENTS

5.1. Experimental Conditions

Detection of a stressed syllable in an input word is carried out based on the maximum likelihood criterion using a word-level score. An input word is matched with a concatenation of stressed/unstressed HMMs. Here, a syllabic transcription of the word, the number of syllables and that of stressed syllables (one in this experiment) of the word are all treated as given. Hence, the number of candidate stress patterns is N for N-syllable input words. Further, in some experiments, syllable boundaries are automatically detected and used explicitly for the stress detection, as shown in Table **1**. The boundary detection is performed in the same manner as in Section **3** except that, for words spoken by Japanese, forced Viterbi alignment is carried out using the concatenation of phoneme HMMs *adapted to Japanese utterances*. In Table **1**, the number of (un)stress HMMs is also shown in the form of an addition of two numbers, which indicate stressed and unstressed HMMs respectively. Position of the stressed HMM in the concatenation which produces the highest word-level score is identified as *stressed*. Figure **1** shows a diagram of the stress detection in the case that the syllable boundaries are explicitly detected and used. In this case, the input word is firstly divided into syllables and then, each of the syllables is matched with its corresponding HMM in candidate stress patterns.

Table 1: Experimental conditions

case	# HMMs	boundaries
a-1	1+1	not detected
a-2	1+1	detected
b-1	8+8 (*str*)	not detected
b-2	8+8 (*str*)	detected
c-1	3+3 (*pos*)	not detected
c-2	3+3 (*pos*)	detected
d-1	18+24	not detected
d-2	18+24	detected

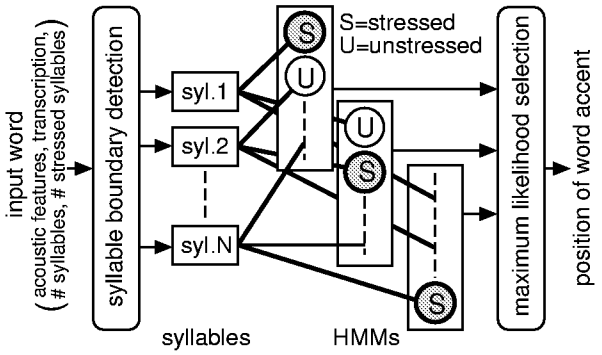

Figure 1: Accent detection using syllable boundaries

5.2. Results and Discussion

As told in Section **3**, each of the testing data of native and Japanese students has its evaluation scores by two English teachers. After normalization of the scores between the teachers, the testing data were categorized into 4 groups separately for native and Japanese utterances. Figures **2** and **3** show the results of each case of the experiments as a function of an average of the normalized scores in a group. Four numbers at the top of each figure indicate the numbers of words belonging to each group. Tables **2** and **3** show averaged detection rates of the testing data of native and Japanese speakers respectively.

For native utterances, comparison between cases **a-1** and **a-2** shows the effectiveness of the explicit use of detected syllable boundaries. Performance increase from cases **a-1** to **b-1** indicates the validity of introducing *structural* information of a syllable. However, there is little difference between cases **a-2** and **b-2**, which implies that the introduction of *structural* information merely results in detecting the syllable boundaries automatically. On the other hand, as for the introduction of *positional* information, not only does it improve the performance from cases **a-1** to **c-1**, but also the information works well even when the syllable boundaries are explicitly used. It can be seen in comparison between cases **a-2** and **c-2**. And the score of case **c-2** exceeds that of case **b-2**. It means that, using the *positional* information, it is possible to characterize other aspects of (un)stressed syllables in the HMMs than those which are already modeled in case **a-2** or **b-2**. In cases **d-1** and **d-2**, since (un)stressed segments are modeled by a unit of phoneme, *structural* information of each syllable is integrated into a concatenation of the models as a matter of course. Therefore, case **d-1** outperforms case **a-1**, but little difference is observed between cases **d-1** and **b-1**. Although this maybe indicates that F_0, power, or duration distribution is *not* phoneme specific,

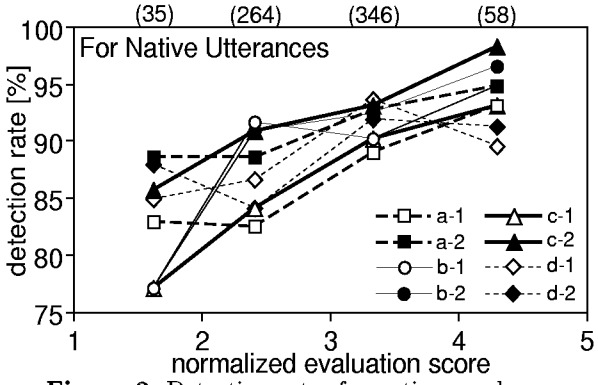

Figure 2: Detection rates for native speakers

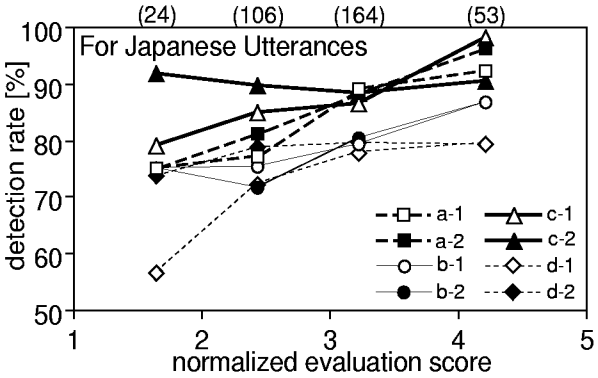

Figure 3: Detection rates for Japanese students

Table 2: Averaged detection rates for native speakers

case	**a-1**	**a-2**	**b-1**	**b-2**	**c-1**	**c-2**	**d-1**	**d-2**
%	86.8	91.4	90.4	91.9	88.7	93.0	90.3	88.9

Table 3: Averaged detection rates for Japanese students

case	**a-1**	**a-2**	**b-1**	**b-2**	**c-1**	**c-2**	**d-1**	**d-2**
%	86.3	87.9	79.7	79.9	87.9	89.5	75.5	80.5

insufficiency of training data cannot be denied. For some models, the number of training data were only around thirty. Further works should be required in this point. In case **d-2** with segment boundaries, the detection rate is decreased unexpectedly. This is considered to be due to less reliability of detected phoneme, not syllable, boundaries as well as data insufficiency.

As for the results of Japanese students, cases **c-1** and **c-2** outperform cases **a-1** and **a-2** respectively as in native speakers. And the maximum detection rate is found in case **c-2** though 3.5 % drop is observed compared to native performance. It indicates the effectiveness of introducing *positional* information into the models and using syllable boundaries detected with HMMs *adapted to Japanese utterances*. However, two major differences can be seen between native and Japanese speakers. Firstly, introducing *structural* information degrades the performance severely, shown in cases **b-1**, **b-2**, **d-1**, and **d-2**. This can be partly explained as following; since Japanese syllables always end with a vowel, Japanese students are inclined to insert an additional vowel after a consonant. This kind of pronunciation habit is supposed to generate the *structural* distortion of syllables. Secondly, while case **c-2** in native speakers shows good performance at every level of the evaluation score, case **c-2** in Japanese students does *not* at higher levels. Cause of this phenomenon is currently investigated. If it is clarified, since case **c-2** shows remarkably good performance at lower levels, the method of case **c-2** is expected to be quite robust in detecting English word accent spoken by Japanese students.

6. CONCLUSION

In this paper, the detection of accent in English words spoken by native and Japanese students were investigated. In the detection experiments, several modifications of acoustic models were examined and, as a result, the use of *positional* information of syllables in a word and adequately detected syllable boundaries was shown to be effective. Although *structural* information was found to be harmful to Japanese utterances, proper adaptation of the (un)stress HMMs is expected to improve the performance as phoneme HMMs for detecting the syllable boundaries of Japanese utterances. Additionally, as future works, the following issues should be focused upon. Phoneme-size (un)stress HMMs trained with a sufficient amount of data, processing of words with secondary accents, which are not dealt with in this study, and quantitative evaluation of strength of word accent, which is referred to as step **ii** in Section **1**. Further, prosodic evaluation on sentence level, namely, intonation, should also be investigated.

REFERENCES

1. S. Hiller *et al.*, "SPELL: An automated system for computer-aided pronunciation teaching," *Speech Communication 13,* pp.463-473 (1993).
2. 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, **J**).
3. A. Cutler *et al.*, "The predominance of strong initial syllable in the English vocabulary," *Computer Speech and Language, 2,* pp.133-142 (1987).
4. 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).
5. N. Ohashi *et al.*, "Pronunciation evaluation of English words spoken by Japanese students based on accent identification using HMMs," *Report of Spring Meet. Acoust. Soc. Jpn.,* pp.117-118 (1997, **J**).
6. S. Takebayashi *et al.*, "A primer of English phonetics (shokyū eigo onsei gaku)," published by Taishūkan shoten (1991, **J**).
7. 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).
8. G. J. Freij *et al.*, "Lexical stress estimation and phonological knowledge," *Computer Speech and Language, 4,* pp.1-15 (1990).
9. P. Dumouchel *et al.*, "Using stress information in large vocabulary speech recognition," *Proc. Montreal Symposium Speech Recognition* McGill Univ., Montreal, pp.73-74 (1986).
10. H. Fujisaki *et al.*, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J.Acoust. Soc. Jpn., 4,* pp.233-242, (1984, **J**).
11. K. Kumpf *et al.*, "Automatic accent classification of foreign accented Australian English speech," *Proc, ICSLP'96* pp.1740-1743, (1996)

J = "in Japanese"