Performance Comparison among HMM, DTW, and Human Abilities in Terms of Identifying Stress Patterns of Word Utterances

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

We have been focusing on applying speech technologies to pronunciation learning. In our previous study[1], a stressed syllable detector was implemented by using stressed syllable HMMs and unstressed ones. And using the detector internally, several systems were implemented[2]. However, their development did not necessarily require the use of HMMs as an acoustic modeling method. In this paper, an HMM-based method, a DTW-based method, and a human strategy only with visual inspection were compared in terms of their performance in judging whether two utterances of a word have the same stress pattern, e.g. *récord* and *recórd*. Here, one utterance was given by a Japanese learner and the other one was done by a native speaker. Experiments showed that HMMs gave us the higher performance than DTW and even human strategies. This result strongly supports the use of HMMs as an acoustic modeling method in the stressed syllable detector development.

1. Introduction

Recently, more and more efforts have been made to apply speech technologies to language learning, especially to assisting pronunciation learning. The authors have been focusing on Japanese manners of generating English word stress. This is because pronunciation habits inevitable to Japanese can be easily found in the stress generation[3]. In our previous studies, several new techniques were proposed to analyze the stress generation manners, namely, the detection of a stressed syllable[1], the estimation of pronunciation habit in terms of the stress generation[2], the abstract visualization of the estimated habit[2], and so forth. These new techniques except the first one were based upon the stressed syllable detection technique, where all the stressed and unstressed syllables were grouped into several tens of syllable classes and each class was separately modeled by HMMs. To build the HMMs, four acoustic parameters of vowel quality, power, pitch and duration were used because these four factors can distinguish between stressed syllables and unstressed ones[4][5]. However, the development of the detector, the estimator, and the visualizer did not necessarily require the use of HMMs as an acoustic modeling/matching method. They can surely be implemented by using another method such as DTW.

Furthermore, we can find even pronunciation CALL softwares where learners are asked to judge whether their utterance and the teacher's one are identical or not in terms of their stress patterns only by inspecting acoustic observations visually presented to them. In this case, learners can be easily assumed *not* to have enough knowledge on acoustics and on the relations between the acoustic observations and differences between stressed syllables and unstressed ones. Therefore, the performance of learners' judging the identity of the two utterances may be rather low.

Based upon these considerations, the performance comparison was carried out among an HMM-based method, a DTW-based method, and a human strategy only with visual inspection. The comparison can certainly be done in the paradigm of identifying a stress pattern given only acoustic observations of an input word. In this paradigm, however, a human subject comes to have to identify a stress pattern without any model utterance. Since this situation can *never* be found in any CALL software, the comparison was designed to be done in terms of the performance in judging whether two utterances of a word have the same stress pattern or not, e.g. *súbject* and *subjéct*. Here, one utterance was given by a Japanese learner and the other one was done by a native speaker.

Section **2.** describes the acoustic modeling of (un)stressed syllables adopted in our previous study. Section **3.** shows the speech sample preparation and Section **4.** outlines the procedures of the experiments. Results and discussions are shown in Section **5.** and Section **6.** concludes this paper.

2. Modeling Stressed Syllables and Unstressed Syllables

The main objective of this paper is to verify the acoustic modeling method adopted in our previous study. Then, this section describes the adopted method.

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. When building syllable models, F_0 for unvoiced segments was required. For these segments, F_0 was estimated by performing the linear interpolation between the preceding and the succeeding voiced segments and the smoothing of the interpolated F_0 curve. After the acoustic analysis, 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 these parameters, duration controlled CDHMMs with six states and four distributions were adopted, where a PDF was comprised of a single Gaussian distribution with a full covariance matrix. The correlation between any two of the above three streams was assumed to be zero in the matrix.

English is estimated to have as many as approximately ten thousand different syllables and this fact led the authors to group the syllables into syllable classes. In our previous studies[1][2], the following several manners of grouping were experimentally examined. And each syllable class was acoustically modeled by the above HMMs.

- Into 2 classes; stressed syllables and unstressed ones. Only the *accentual* attribute is considered. ⟶ **simple**
- Into 6 classes; S_H, S_T, and S_O separately for stressed syllables and unstressed ones, where S_H and S_T denote a syllable at the head and one at the tail of a word respectively, and S_O indicates a syllable at the other parts of the word. Here, the *accentual* and *positional* attributes of the syllable are introduced into the HMMs. ⟶ **pos**
- Into 16 classes; V_S, CV_S, V_SC, CV_SC, V_L, CV_L, V_LC, and CV_LC separately for stressed syllables and unstressed ones, where V_S/V_L represents a short/long vowel and C means a sequence of consonants whose length is more than zero. In this grouping, the *accentual* and *structural* attributes are integrated into the HMMs. ⟶ **str**
- Into 48 classes; all the above three attributes are considered. It follows that the models are built separately for each combination of their accentuation, their position in the word, and their syllabic structure. ⟶ **pos_str**
- Into 80 classes; the *contextual* attribute is additionally introduced to further refine grouping **pos_str**. Here, four labels — stressed / unstressed / begin_of_word / end_of_word — are prepared as values of the left-hand or right-hand contextual attribute. ⟶ **context**

3. Preparation of Speech Samples for the Comparison Experiments

The comparison experiments required two kinds of words spoken by Japanese; 1) word utterances whose stress patterns are identical to those of the same words spoken by a native speaker, and 2) word utterances whose patterns are different from those of the same words spoken by a native speaker. These utterances could be easily obtained by using words which can have different stress patterns, such as *record*, *object*, and so on. If these words only were used, however, the scale of the experiments should be rather small. Therefore, we attempted to collect word utterances of both types using a larger vocabulary. In this case, it should be noted that the collection should *not* be done by forcing speakers speak words with different stress patterns from those which they thought were correct for the words. This is because the forced utterance may cause unnaturalness in the speaking. Then, the collection was done so that the following conditions were satisfied;

- Japanese speakers spoke words with stress patterns which they thought were correct for the words.
- Half of the obtained word samples were uttered with *lexically* correct stress patterns and the other were not.

To realize the speech sample collection, a test was carried out beforehand where eight Japanese speakers, **A** to **H**, were required to locate the stressed syllable in each word of an English word set. Here, the number of words of the set was one hundred. Using the results of the test, a set of sixty-five words were selected so that, for every speaker, the indicated locations of the stressed syllables of only about half words of the selected set were lexically correct. Although the selected set was common to all the speakers, the indicated locations of the stressed syllables were dependent on the individual speakers. For the speech sample preparation, the eight Japanese, **A** to **H**, were asked to speak all the words of the selected set with the stress patterns which they answered in the test. And the spoken words were recorded with a DAT recorder and analyzed as in Section **2.**. As for the word samples spoken by a native speaker, ATR English word database was used, which contains speech samples of the 5000 most frequent words.

4. Procedures of the Experiments

4.1. Judging experiments with the HMM-based method

For each of two utterances of the same word, one spoken by a Japanese and the other spoken by a native speaker, the stressed syllable was detected by the method proposed in our previous study. Here, the **pos** HMMs trained with ATR English database of a different speaker from the native speaker used in Section **3.** were adopted for the stress detection. The detection procedure is schematically shown in **Figure 1**. If the positions of the detected stressed syllables were the same between the two utterances, the two words were judged to have the same stress pattern. It should be noted that the above requirement for the same stress pattern does not always request that the detection should be done correctly. Even if the detection did not work right, the two word utterances were judged to have the same stress pattern when the positions of the detected stressed syllables were the same between the two.

4.2. Judging experiments with the DP-based method

For each of the two utterances, starting and ending frames of the word segment were firstly detected automatically. This detection was also done in the other experiments. After that, the two segmented words were matched with each other based upon DP-matching. Parameter vectors used here were the same as those in the previous section, namely, a concatenation of cepstrum-related, power-related, and pitch-related parameters. The judgment in respect to the

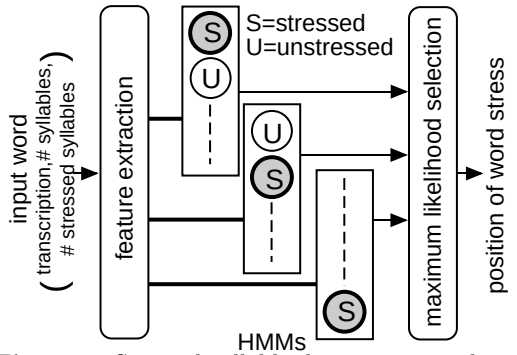

Figure 1: Stressed syllable detection procedure

stress pattern was made by adequately setting a threshold to the averaged distance over frames between the two words. And if the averaged distance was smaller than the threshold, the two words were judged to have the same stress pattern. In this case, however, a problem could occur. A parameter vector used in the experiments was composed of three different acoustic observations, i.e. cepstrum, power, and pitch. This might result in a rather large difference of the range of the local distance among individual elements. Therefore, the local distance of an element of a vector was calculated after the normalization based upon the variance of the element. And the local distance, LD, of a full vector was calculated as in the following equation,

$$LD(i,j) = \sum_{d=1}^{12} \frac{\{v_i(d) - w_j(d)\}^2}{\sigma(d)^2}, \quad (1)$$

where i and j are frame indexes of two input vectors v_i and w_j and (i,j) means a point in the DP path. d denotes the dimension of the vector and $\sigma(d)^2$ means the variance of the d-th dimensional element of the vector. The variance was calculated by using the whole training data for HMMs.

In addition to the simple definition of LD, its extension was experimentally examined and it was defined as

$$LD(i,j) = \rho_1 \sum_{d=1}^{8} \frac{\{v_i(d) - w_j(d)\}^2}{\sigma(d)^2} + \rho_2 \sum_{d=9}^{10} \frac{\{v_i(d) - w_j(d)\}^2}{\sigma(d)^2} + \rho_3 \sum_{d=11}^{12} \frac{\{v_i(d) - w_j(d)\}^2}{\sigma(d)^2}, \quad (2)$$

$$\text{where } \sum_{i=1}^{3} \rho_i = 3.0 \quad (\rho_i \geq 0.0). \quad (3)$$

In this equation, three weights, $\{\rho_1, \rho_2, \rho_3\}$, are assigned to three sub-scores, which represent cepstrum, power, and pitch terms. And Equation (**1**) can be interpreted as one special case of Equation (**2**), i.e. $\rho_i = 1.0$ for each i. This extension was derived due to the following two considerations. The normalization of the distance range of individual elements does not always improve the performance. And the adequate emphasis/deemphasis of a specific feature should be effective, although the optimal control of weights must be speaker-dependent.

For either definition of the local distance, the DP path could be drawn in two different manners. One is an ordinary manner, which temporally aligns the input two words using vectors comprised of cepstrum, power, and pitch sub-vectors (*ordinary alignment*). The other is as follows. Firstly, the DP path was drawn by conducting the DP-matching between the two words with one to ten dimensions of LPC mel cepstrum coefficients and their derivatives, namely, with segmental features (*segmental alignment*). After that, the summation of the local distances, Equation (**1**) or (**2**), was done along the obtained DP path. Then, the judgment of the stress patterns' identity was done by using the averaged distance over frames.

4.3. Judging experiments with the human strategy with visual inspection

Subjects were visually provided with acoustic observations of the two word utterances. The visualized acoustic features were power and pitch patterns only. Here, spectrum patterns were not presented although they were used as one of the acoustic parameters both in the HMM and DP-based methods. This is because we could not find any commercial software in Japan which showed spectrum patterns of learners' utterances, although we made a detailed survey of eighteen CALL softwares available in the Japanese market. In other words, this experiment was designed so that the situation of using the current CALL softwares was realized. The presentation of power and pitch patterns was conducted in three different manners. One was just showing power and pitch patterns without any alignment. Another one was displaying them after the ordinary alignment between the two utterances. The last one was the same as the previous one except for acoustic features used in the alignment. Here, the segmental alignment was performed. **Figure 2** shows an example of the power and pitch patterns with the ordinary alignment. In the experiments, audio feedback of the utterances was *not* done because speech signals were *not* used directly as one of the acoustic features in the HMM or DP-based method.

Subjects of this experiment were eight Japanese, **a** to **h**, a part of whom participated in the recording of English words in Section **3.**. To all the subjects, the same set of word pairs, Japanese and native utterances, were presented with their power and pitch patterns. Here, the set was comprised of 130 word pairs, half pairs of which were spoken by **A** and a native speaker, and the other pairs of which were by **B** and the same native speaker. Although **A** was referred to as **a** in this experiment, he could not recognize that he was looking at his own utterances because there was no audio feedback during the experiments.

4.4. Quantitative measurement of judging performance

After the three experiments, *judging performance*, JP, was calculated for each method by the following formula.

$$JP = \frac{\frac{n_s}{N_s} + \frac{n_d}{N_d}}{2} \times 100 \quad [\%] \quad (4)$$

In the formula, while N_s indicates the number of word pairs which had the same stress pattern, n_s means the number of word pairs which were judged to have the same pattern

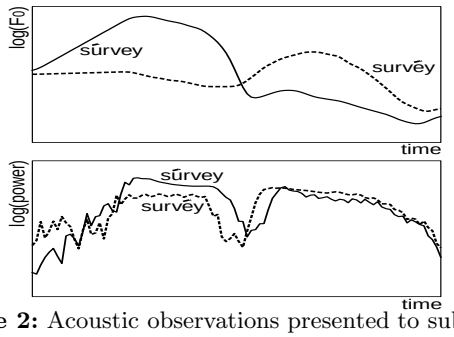

Figure 2: Acoustic observations presented to subjects

Table 1: JP using the HMM-based method [%]

speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	avg.
JP	76.9	83.0	83.5	88.1	68.8	86.2	80.0	89.3	82.0

Table 2: JP using the DP-based method [%]

With the *simple LD* and the *ordinary* alignment.

speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	avg.
JP	62.1	59.3	63.3	56.7	62.6	71.3	50.1	62.0	60.9

With the *extended LD* and the *ordinary* alignment.

speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	avg.
JP	69.3	63.9	67.3	66.7	64.3	77.7	57.9	71.0	67.3
ρ_1	0.5	0.0	0.5	0.0	0.5	0.0	1.5	0.0	
ρ_2	1.5	0.5	2.5	3.0	0.5	1.5	1.5	1.5	
ρ_3	1.0	2.5	0.0	0.0	2.0	1.5	0.0	1.5	

With the *simple LD* and the *segmental* alignment.

speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	avg.
JP	61.9	74.1	64.5	56.1	50.0	88.9	58.4	60.6	64.3

With the *extended LD* and the *segmental* alignment.

speaker	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	avg.
JP	73.8	75.8	67.3	71.8	66.0	92.2	62.1	65.1	71.8
ρ_1	0.0	1.5	1.5	0.0	0.0	0.5	0.5	0.0	
ρ_2	1.5	0.5	1.0	3.0	3.0	2.5	2.5	2.0	
ρ_3	1.5	1.0	0.5	0.0	0.0	0.0	0.0	1.0	

Table 3: JP using the human strategy with inspection [%]

With no alignment.

subject	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	avg.
speaker **A**	73.8	79.3	79.8	64.5	79.5	75.7	49.8	74.5	72.1
speaker **B**	78.8	69.2	69.2	69.4	68.9	73.7	75.2	80.5	73.1

With the *ordinary* alignment.

subject	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	avg.
speaker **A**	66.2	71.9	76.2	61.7	71.7	62.9	51.9	78.1	67.6
speaker **B**	73.9	76.9	73.1	62.1	69.0	57.9	81.1	78.6	71.6

With the *segmental* alignment.

subject	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	avg.
speaker **A**	73.1	79.8	83.6	71.7	77.4	74.0	59.8	89.0	76.1
speaker **B**	79.7	66.2	75.2	69.0	59.6	60.8	72.0	79.7	70.3

in the experiments. N_d and n_d are the numbers of word pairs for the case of different stress patterns.

5. Results and discussions

Table 1 shows judging performance (JP) of the HMM-based method separately for each of eight speakers, **A** to **H**. And the averaged performance is 82.0%.

In **Table 2**, JP values of the DP-based method are listed for each speaker in various conditions; two definitions of LD $\times$ two manners of temporal alignment. In the tables of extended LD, the quasi-optimal combinations of weights, $\{\rho_1, \rho_2, \rho_3\}$, are also shown for individual speakers. In the experiment, these combinations were estimated by a greedy method, where a triangular plane in (ρ_1, ρ_2, ρ_3) space characterized by Equation (**3**) was firstly quantized to give several tens of representative points on the triangle, and then the point which maximized JP was selected as the quasi-optimal weight combination. Although the weight optimization gives us the highest performance, this performance cannot be realized for a *new* speaker because his/her optimized combination is unknown. Both of the use of extended LD and that of segmental alignment increase JP. However, the averaged JP in every condition is lower than the averaged JP of the HMM-based method.

Table 3 shows JP of the human strategy with inspection in three conditions of no, ordinary, and segmental alignments. Unlike JP of the DP-based method, the use of temporal alignment does not necessarily increase the performance. This is considered to be because subjects unconsciously aligned the visualized patterns of one utterance with those of the other while comparing the two utterances. Therefore, JP is not always decreased in the case of 'no alignment'. Since this experiment used word samples spoken by **A** and **B** only, each JP value in **Table 3** should be compared with JP of **A** and that of **B** shown in **Table 1**, which are 76.9% and 83.0% respectively. Although, in some cases, JP values in **Table 3** exceed the corresponding ones in **Table 1**, all the averaged JP values of each speaker in **Table 3** are lower than the corresponding values in **Table 1**. This result implies that subjects without any special training do not have enough knowledge on how to evaluate differences between the two utterances and on how to integrate difference between the two power patterns and that between the two pitch patterns. And HMMs are considered to acquire this knowledge through their training procedures with a spoken word database.

6. Conclusions

In this paper, to verify the acoustic modeling/matching method adopted in our previous study, the performance in judging whether two utterances of a word have the same stress pattern or not was examined among an HMM-based method, a DP-based method, and a human strategy with visual inspection. To increase the reliability of experimental results, word speech samples were carefully prepared, where a test was carried out beforehand to estimate speakers' knowledge on English vocabulary. Judging experiments showed that the HMM-based method gave the higher performance than the DP-based method and even the human strategy. This result strongly indicates the validity of using HMMs as an acoustic modeling/matching method in the stressed syllable detector development.

References

1. N. Minematsu *et al.*, "Automatic detection of accent in English words spoken by Japanese students," Proc. EUROSPEECH'97, pp.701–704 (1997).
2. N. Minematsu *et al.*, "Prosodic evaluation of English words spoken by Japanese based upon estimating their pronunciation habits," Proc. ICSP, pp.439–444 (1999).
3. 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).
4. S. Hiller *et al.*, "SPELL: An automated system for computer-aided pronunciation teaching," Speech Communication, vol.13, pp.463–473 (1993).
5. A. Ljolje *et al.*, "Recognition of isolated prosodic patterns using Hidden Markov Models," Computer Speech and Language, vol.2, pp.27–33 (1987).