ACOUSTIC MODELING OF SENTENCE STRESS AND ITS DETECTION FOR LEARNING ENGLISH RHYTHM

Nobuaki MINEMATSU† *Satoshi KOBASHIKAWA*† *Keikichi HIROSE*† *Donna ERICKSON*‡

†University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
‡Gifu City Women's College
7-1, Hitoichibakita-machi, Gifu-shi, Gifu 501-0192, JAPAN
{mine, kobashi, hirose}@gavo.t.u-tokyo.ac.jp erickson@gifu-cwc.ac.jp

ABSTRACT

This paper proposes a new technique for acoustic modeling of stressed/unstressed syllables in sentence utterances of American English. Here, relative differences of acoustic features between two consecutive syllables characterizing "stressed" or "unstressed" were introduced into the HMM-based acoustic modeling. This is because syllables can be identified as stressed or unstressed only after comparing them with their neighboring syllables. For training syllable HMMs, speech samples were recorded by ourselves because we could not find any database which can be directly used for this modeling. The fourth author put multi-level stress marks (syllable magnitude) on individual syllables of a given sentence set, which was done according to guidelines for teaching English rhythm to non-native speakers of English, proposed and used in class by the fourth author. After the stress mark assignment, the sentences were uttered by her and recorded to be used for the HMM-based modeling. Experiments showed that stress/unstress identification errors were reduced by about 25% in comparison to the modeling technique without the relative differences.

1. INTRODUCTION

Recent advances in speech recognition techniques make it possible to develop CALL systems especially for pronunciation learning. Since the speech recognition techniques have been basically devised only to identify phonemes, acoustic features irrelevant to the identification are often discarded. 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 support only the phonemic aspect of pronunciation learning.

We can easily find many studies which emphasize the importance of learning pronunciation in terms of prosody. English words with wrong stress patterns tend to be more difficult for native speakers to accept than those with wrong phonemic features[1]. Native speakers perceive word stress as essential information for identifying isolated word utterances[1]. The assignment of incorrect stress patterns to words degrades the transmission of segmental information[2]. These works clearly indicate that pronunciation learning in terms of prosody is as important as in terms of phonemic features.

In our previous studies, a technique was proposed to model English lexical stress generated in isolated word utterances[3, 4].

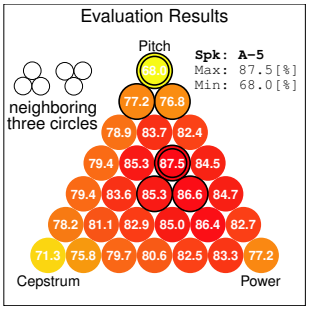

Fig. 1. An example of triangular representation of pitch, power, and vowel quality of learners' utterances.

Syllables were clustered into several tens of syllable classes according to positional attributes of the syllable in a word and structural attributes of the syllable. Four acoustic features of power, pitch, duration, and vowel quality, which were said to be affected by accentual attributes of syllables, were integrated into HMM-based modeling of each of the syllable classes. Using the modeling technique, a stressed syllable detector was developed. A visualizer was then built, based on Japanese-specific manners of controlling the four acoustic factors when generating stressed syllables. The visualization was implemented using the abstract and integrated representation of acoustic information. Direct presentation of acoustics observed in learners' utterances are not always adequate because they rarely have enough knowledge to understand the presented acoustics. Furthermore, separate presentation of the four acoustic factors is not necessarily good because the learners don't know how to integrate the presented factors. Our proposed visualizer solved the two problems simultaneously and experiments showed that the visualized patterns had quite high correlation with pronunciation proficiency scores rated by human English teachers. **Figure. 1** shows an example of the abstract and integrated representation of acoustic information, called a "triangular representation."

In this paper, following our previous studies, a sentence stress detector was developed where research focus was placed upon selecting acoustic features adequate especially for modeling stressed syllables in sentence utterances. In sentence utterances, some of the four acoustic factors, power, pitch, duration, and vowel quality, were affected rather largely by linguistic events other than sentence stress. A typical example is intonation, by which pitch patterns can

be changed easily. In this paper, we investigated relative acoustic differences between consecutive syllables in order to look at local variations of the acoustic features. Experiments showed high validity of the proposed modeling method of sentence stress.

2. PREPARATION OF SENTENCE SPEECH SAMPLES WITH STRESS LABELS

2.1. Strategy for collecting speech samples with stress labels

To build HMMs for stressed syllables and unstressed ones, a speech database with sentence stress labels was required. However, most of the speech databases were developed to be used as training/testing samples of speech recognizers. This means that available databases do not contain labels of sentence stress. These labels can be assigned to any of the available speech databases by asking phoneticians to do the stress marking. However, this would result in large variation among labelers. One solution to reduce labeler-dependency and speaker-dependency is to ask many phoneticians to label a large speech database. But it is evident that this solution is quite costly. In this work, we adopted another strategy. We asked that the stress markings be done by an English teacher (the fourth author) who had developed a method of teaching English rhythm to non-native speakers of English and who also was a speech researcher with good knowledge of prosody. After the stress marking to a given sentence set, all the sentences were uttered by her and recorded for the modeling. After the recording, all the speech samples were checked by her listening to whether the individual syllables were adequately produced according to the stress labels which she assigned before the recording. The recorded speech samples were divided into two sets for training and testing (Set-A0-a/b). Additional speech samples were prepared for testing, which were from CDROMs of an English courseware (Set-A1) and TIMIT database (Set-A2). Further, English sentences spoken by Japanese students were collected (Set-J). **Table. 1** shows speech samples prepared. The stress marking for Set-A1/2 and Set-J was also done by the fourth author by listening to the samples.

This strategy of preparing speech samples has an evident defect in that HMMs built only based on a single speaker will surely have large speaker-dependency. This problem can be solved by collecting speech samples spoken by others, which is often done in the speech recognition community. The main research focus in this paper is whether HMMs built only with *her* speech samples can automatically identify stressed syllables in *her* sentence speech as correctly as *she* can identify them by listening. In other words, the performance of the proposed method has to be comparable at least with her own ability of identification by listening. Otherwise, the proposed modeling technique has some clear drawbacks for sentence stress detection.

Table 1. English speech samples used in the experiments

1) For training

set	content
Set-A0-a	592 sentences spoken by the fourth author

2) For testing

set	content
Set-A0-b	120 sentences spoken by the fourth author
Set-A1	75 sentences spoken by a male native American English speaker
Set-A2	90 sentences spoken by 15 male and 15 female native American English speakers
Set-J	125 sentences spoken by 5 male and 5 female Japanese students

2.2. Sentence stress assignment done before the recording

The kinds of labels assigned to individual syllables, done by the fourth author, are described as follows. They are based on a practical approach she developed for approximating syllable magnitudes when she taught ESL at The Ohio State University[5]. An assumption is that English rhythm results from different weightings assigned to syllables, based on "levels" of sentence structure. The levels are simplified to include only four levels: (1) syllable, (2) word, (3) phrase, and (4) sentence. **Table. 2** show some examples of sentences with the four-level stress marking. Numbers correspond to levels of syllable magnitude, which range from 1 to 4. Here are some basic principles for the assignment. Each syllable gets level 1 at the minimum. The stressed syllable of each polysyllabic word gets level 2 at the minimum. The stressed syllables of the final content word in a phrase and that in a sentence often get levels 3 at the minimum and 4, respectively. It should be noted that this stress assignment method is considered to give ESL students just one reference stress pattern and it is not claimed that only this pattern is acceptable to native speakers of English. In her works, she tentatively proposed several modifications for the principles. For example, more-than-four level marks and shift of the largest stress were examined for emotional and/or emphasized speech.

In the rest of this paper, unstressed syllables were defined as those of level 1 and stressed syllables were as those of level 2 or more. In all the experiments, binary identification of stressed syllables and unstressed ones were carried out. Separate modeling of syllables of each level is left as a future project.

3. MODELING OF STRESSED/UNSTRESSED SYLLABLES IN SENTENCE UTTERANCES

English is estimated to have as many as approximately ten thousand different kinds of syllables. Therefore in this paper, English syllables were grouped into syllable classes in terms of their accentual, positional, and/or structural attributes to be modeled by

Table 2. Examples of stress assignment to individual syllables

1 1 4 I'm amused.
1 1 3 | 1 1 4 I'm amused | by the man
1 1 3 | 1 1 3 | 1 1 4 I'm amused | by the man | and his jokes.
1 1 3 | 1 1 3 | 1 1 2 1 4 I'm amused | by the man | and his funny jokes.
2 1 2 4 1 Betty cooks breakfast.
2 1 2 1 2 4 1 Betty often cooks breakfast.
2 1 2 1 2 3 1 | 1 2 4 1 Betty often cooks breakfast | before seven.

1, 2, 3, and 4 = levels of stress, | = phrase break

HMM. The grouping methods used here are described below.

SIMPLE (2 classes) : stressed and unstressed syllables. Only the *accentual* attribute of the syllable is considered.
POS (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 tail of a phrase, and S_O indicates a syllable at the other parts. In this case, the *accentual* and *positional* attributes of the syllable in the phrase are introduced into the HMMs.
STR (24 classes) : V_X, CV_X, V_XC, CV_XC (X=L, S, D) separately for stressed and unstressed syllables, where V_S, V_L, and V_D represent a short vowel, a long vowel, and a diphthong respectively and C is a sequence of consonants. In this grouping, the *accentual* and *structural* attributes of syllables are integrated.
POS_STR (72 classes) : integration of the above two groupings.

To model each of the above syllable classes, four acoustic features of power, pitch, duration, and vowel quality were used. Speech samples were digitized with 12 kHz & 16 bit sampling and the 14-th order LPC analysis was carried out using 21.3 ms frame length and 8.0 ms frame rate to calculate LPC mel cepstrums. *F0* and power were also extracted with the same rate and, after being transformed to the logarithmic scale, they were shifted to have zero as mean values over all the utterances in our baseline method and over local segments in our proposed method. Detailed description of the proposed normalization will be found in the following section. *F0* values for unvoiced segments were required for the modeling and were estimated by non-linear interpolation between the preceding voiced segment and the succeeding one. After the analysis, 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) *F0* and its derivative. It should be noted that cepstrum coefficients were calculated after CMN (Cepstrum Mean Normalization). Using the streams, each of the syllable classes was acoustically modeled by using duration controlled continuous density HMMs.

4. EXPERIMENTS OF DETECTING SENTENCE STRESS

4.1. Detection with the conventional modeling technique

As the baseline performance, we experimentally evaluated the stress modeling with the previously proposed technique[3], where *F0* and power were normalized over all the utterances. Before the detection, input speech was segmented syllable by syllable. This segmentation was done with phoneme-level forced alignment and an automatic syllabification tool which took an arbitrary phoneme sequence as text input and estimated syllable boundaries in the sequence[6]. By comparing likelihood score as stressed and that as unstressed for each segmented syllable, the stress detection was carried out. Likelihood score calculation was done by matching the acoustic features of the segmented syllable with the corresponding HMM. In the process of detecting stressed syllables, we can introduce a word-level constraint that each polysyllabic word must have only one stressed syllable. This constraint may not be good for detecting the stressed syllable from utterances of students with poor pronunciation proficiency, and therefore, all the detection results will be shown separately for two cases, namely, with/without that constraint. **Figure. 2** shows the baseline performance of the stress detection from sentence utterances.

In the speaker-closed condition (Set-A0-b), the finer models gave us the higher performance and the constraint was shown to improve the performance. On the other hand, for other native American English speakers, while the constraint brought about the improvement, the modeling with structural attributes didn't work well. This is because the models were built in a speaker-specific mode; this can be solved by collecting speech samples of other speakers. As for Set-J, the improvement by the positional attribute of syllables was shown but the absolute performance of the detection was still low. To improve the performance, some other methods will be devised in section **4.4** for Japanese utterances.

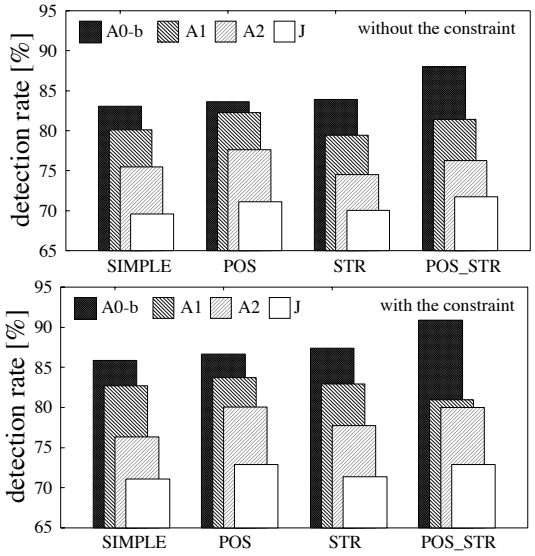

Fig. 2. The baseline performance obtained with and without the word-level constraint

4.2. Detection with relative differences between syllables

Syllables can be perceived as stressed when they have larger sonority than their neighboring syllables. This means that the detection of stressed/unstressed syllables should be done by looking at relative differences between consecutive syllables, namely, local variation of the parameters. In the previous section, however, the local variation was not captured directly because the normalization was done over all the utterances. The immediate parameterization of the local variation will give us another benefit. Some of the four acoustic features characterizing the accentual attribute of the syllable can be changed by other linguistic events. One typical example is intonation, which can result easily in large *F0* changes. For the stress detection, it was desirable to delete *F0* variation caused by intonation. This intonation-related variation often draws a global pattern compared to the sentence stress-related variation. And therefore, the immediate use of the local variation will ignore the undesired and global variation of the parameters.

In this work, we examined three methods for the local normalization of log-*F0* and log-power. In the first method, the average values over the target syllable and its preceding one were obtained and the normalization was done so that the average values were zero. The normalization could also be done with the target and the succeeding syllables. In the second method, the average values over only the preceding syllable were used for the normalization. In this case, the normalization was possible with the succeeding syllable only. In the first and second methods, the average values over a certain speech segment were shifted to zero. In the third

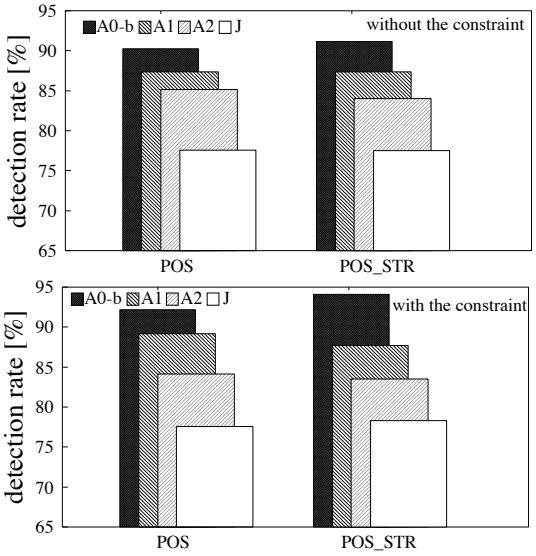

Fig. 3. Stress detection performance with the local normalization

method, however, log-F_0 and log-power at the starting point of the target syllable were shifted to zero. A similar normalization was also possible by shifting the parameters at the ending point.

In each method of the local normalization, two kinds of likelihood scores could be generated: one related to the preceding syllable or the starting point and the other, to the succeeding syllable or the ending point. Another likelihood score could be used here, which is the score calculated in section **4.1**. The detection experiments in this section were carried out by using the two likelihood scores with the local normalization and another score without it. The integration of these three scores was done by summation of them with adequate weightings.

Experiments showed that the performance was improved in every method, irrespective of the speaker group. Due to the limitation of space, the best performance was shown in **Figure. 3** with/without the word-level constraint, which was obtained in the third method. Clearly shown in the figure, the performance was improved in every speaker group and approximately 94% detection rate was obtained in Set-A0-b. It should be noted that a large improvement was observed in **POS** modeling. As shown in section **4.1**, **POS_STR** modeling tended to be speaker-dependent. Large improvement in **POS** modeling was helpful when building the HMMs with a limited amount of training data.

4.3. Performance comparison with human detection ability

As told in section **2.1**, the main focus of this study is to investigate whether the speaker-specific model built with the proposed technique can be comparable with human performance in the stress detection. Listening experiments were carried out, where sets of a three-syllable-sequence were presented through headphones to the fourth author. The sequences were segmented from *her* speech samples so that every segmented sequence had a stressed syllable and an unstressed syllable at least and no sequence comprised a single word. This is because the lexical information should not be used in the listening. The total number of the presented sequences was 190. The task was to judge whether the central syllable was stressed or not. The reason for a three-syllable-sequence presentation was that the proposed technique used the acoustic features related to the neighboring syllables. One-syllable presentation was tentatively carried out. In this case, however, the judgment was dependent on the volume level when presenting the syllables, and therefore the obtained results were considered to be less reliable. As a result, the human detection performance was 95.8 % and the machine performance was also 95.8 % for the 190 sequences with the proposed local normalization, which clearly shows the proposed method can detect completely as well as humans can. It is interesting that, while the performance is the same, errors can be found in different sequences between human and machine.

4.4. Detection only with acoustic features of vowels

In the previous sections, the detection performance was quite low in Japanese speech. This is partly because of the speakers' poor pronunciation ability. The stress labels were assigned to the Japanese samples by the fourth author's listening and it is desirable that this assignment or judgment can be simulated. To improve the simulation performance, the modeling of stress/unstress with only acoustic features of vowels was examined. This is because Japanese students tend not to produce adequate syllable structure in their speech; the stress/unstress modeling with respect only to vowels is found in [7]. Experiments showed about 15 % detection error reduction in Japanese speech while a slight increase of the errors was observed in native American English speech. The vowel-based detection should be carefully treated because English rhythm is composed of syllable units in contrast to Japanese rhythm, which is composed of mora units, often smaller than syllables. However, to improve the stress detection for Japanese speech, the vowel-based detection was shown to be effective.

5. CONCLUSION

In this paper, technical investigations were done mainly in order to correctly detect the stressed syllables from sentence utterances. Here, relative differences between consecutive syllables were introduced to acoustically model stressed and unstressed syllables. Experiments showed the proposed technique had exactly the same performance of a human English teacher with good knowledge of English rhythm. Further, to improve the performance for utterances spoken by Japanese, vowel-based detection was examined, which showed validity only for speech samples of Japanese.

6. 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] N. Minematsu and S. Nakagawa, "Visualization of pronunciation habits based upon abstract representation of acoustic observations," Proc. InSTIL'2000, pp.130–137 (2000).

[4] N. Minematsu and S. Nakagawa, "Instantaneous estimation of prosodic pronunciation habits for interactive instructions to learn English pronunciation," Proc. ICSLP'2000, vol.3, pp.191–194 (2000).

[5] D. Erickson, "Jaw movement and rhythm in English dialogues," Technical Report of Acoust. Soc. Jpn., H-98-59, pp.1–8 (1998).

[6] http://www.nist.gov/speech/tools/tsylb2-11tarZ.htm

[7] C. Wang and S. Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the Jupiter domain," Proc. EUROSPEECH'2001, vol.4, pp.2761–2764 (2001)