MODELING AND AUTOMATIC DETECTION OF ENGLISH SENTENCE STRESS FOR COMPUTER-ASSISTED ENGLISH PROSODY LEARNING SYSTEM

*Kazunori Imoto** *Yasushi Tsubota** *Antoine Raux** *Tatsuya Kawahara** *Masatake Dantsuji*[†]

* School of Informatics, Kyoto University,
[†] Center for Information and Multimedia Studies, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

We address sentence-level stress detection of English for Computer-Assisted Language Learning (CALL) by Japanese students. Stress models are set up by considering syllable structure and position of the syllable in a phrase, which will provide diagnostic information for students. We also propose a two-stage recognition method that first detects the presence of stress and then identifies the stress level using different weighted combinations of acoustic features. The modeling is coherent with conventional linguistic observations. The method achieves stress recognition rate of 95.1% for native and 84.1% for Japanese speakers.

1. INTRODUCTION

Computer-Assisted Language Learning (CALL) has drawn increasing attention thanks to the progress of speech technologies[1][2]. CALL systems provide us with easier repetitive training as well as multi-media dialogue environment[3]. While many CALL systems have been developed focusing on segmental pronunciation and dialogue training, very few deal with evaluation and instruction on prosodic features, namely stress, rhythm and intonation, which play an important role in human communication.

To detect erroneous prosodic patterns, it is significant to take into account that students are often affected by their native language. For example, Kawai et al.[2] proposed a method to detect segmental errors in Japanese English speech using both Japanese and English phoneme models. Minematsu et al.[4] made use of the acoustic difference between a non-native utterance and a native model to generate diagnostic feedback with regard to the stress in an isolated word. However, the method cannot be easily applied to sentence utterances because specific features that affect sentence-level prosody such as intonational phrasing and unstressing of function words need to be considered.

In this paper, we address a CALL system focusing on sentence stress of English by Japanese students. We first investigate factors affecting Englsh sentence stress and errors by Japanese students, and then set up classifications of stressed syllables. The modeling differs from previous studies that deal with native speech and/or isolated words[5][6]. Based on the classifications, we construct HMMs (Hidden Markov Models) to detect the sentence stress and to identify the cause of errors. We also propose a two-stage method, which first judges whether the syllable is stressed and then identifies its stress level, by using different weights of acoustic features estimated by discriminant analysis.

2. MODELING OF ENGLISH SENTENCE STRESS BY JAPANESE STUDENTS

2.1. Characteristics of English Sentence Stress

In isolated words, the position of stressed syllables is fixed. In the case of sentence utterances, stress patterns are affected by component words' context. Typically, content words such as verbs and nouns are stressed whereas function words such as articles and prepositions are not[7]. But many other factors affect the sentence stress and several patterns are often possible. In a CALL system, use of a skit constrains the sentences with some context and enables prediction of a correct stress pattern.

In English, stressed syllables are characterized by not only power level, but also pitch, duration and vowel quality[8]. However, pitch in natural conversation rises rapidly at the beginning of each phrase unit and falls gradually, giving a complex influence on the sentence stress.

2.2. Stress Errors by Japanese Students

Next, the causes of errors by Japanese students are addressed.

(1) Difference between stress and pitch accent

In Japanese, important words are emphasized by the change of pitch. Therefore, Japanese students tend to mark stressed syllables using only pitch instead of whole set of acoustic features that characterize stress in English. This results in perceived errors.

(2) Incorrect syllable structure

English has a large number of possible syllable structures from a single vowel (V) to syllables including as many as seven consonants. By contrast, Japanese syllables are basically limited to V and CV[1]. As a consequence, Japanese students often insert vowels in pronouncing English syllables whose structures do not exist in Japanese. This deformation of syllable structure can further imply stress errors. For example, Japanese students tend to pronounce the word "strike" as /s-u-t-o-r-ay-k-u/ and put stress on the added vowels instead of the main vowel /ay/.

(3) Incorrect phrasing

When pronouncing complex sentences, non-native speakers may divide them into phrase units that do not match the sentences' syntactic structure. Pitch movement implied by improper phrasing leads to stressing of syllables at inappropriate positions.

2.3. Classifications of Stressed Syllables

Based on the observation, we present three classifications of stressed syllables. Their combinations yield different models.

(1) Classification by stress level (base)

We adopt a 3-level sentence stress system. Primary stressed syllables (PS) are syllables that carry the major pitch change in a tonal group (=phrase). Hence, there is only one PS in each phrase, usually placed on the word having the most important piece of information. Secondary stressed syllables (SS) are all other stressed syllables. Finally, non-stressed syllables (NS) are syllables that do not bear any mark of stress.

(2) Classification by syllable structure (syl)

Syllable structure and stress are correlated such that complex structures have larger probability to be stressed[9]. We classify syllables into four categories: V, CV, VC, CVC. We also classify vowels into four types: schwa (Vx), short vowel (Vs), long vowel (Vl), and diphthong (Vd). Thus, combination of these two factors makes 16 categories of syllables.

(3) Classification by position in phrase (pos)

Since pitch movement in the initial and end part of a phrase behaves differently, the prosody pattern is different depending on the position of the syllable in the phrase. Thus, we also classify syllables into three types from the viewpoint of the position in a phrase: head (H), middle (M), and tail (T).

Based on these categories, we set up models for three stress levels with sixteen syllable structures and three positions in a phrase. Thus, in the most detailed case, there are 144(=3*16*3) HMMs.

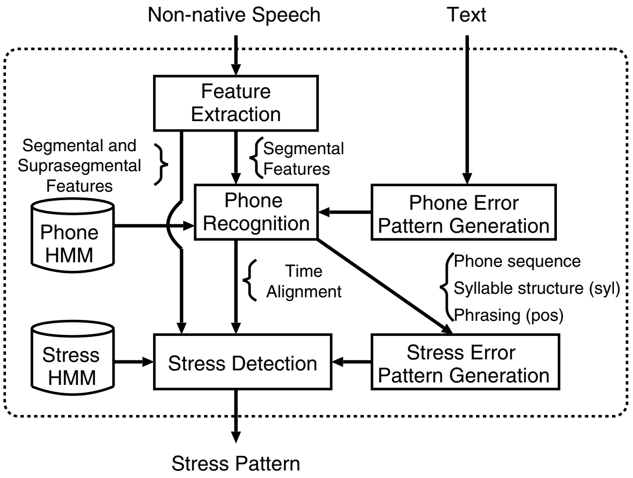

Fig. 1. Flow of detection of sentence stress

3. AUTOMATIC DETECTION OF SENTENCE STRESS

An overview of sentence stress detection is depicted in Figure 1. The process consists of three parts, which are explained in the following subsections.

3.1. Speech Analysis and Feature Extraction

Acoustic features take into account pitch ($\log(F_0)$), power ($\log(power)$) and spectral (MFCC: Mel-Frequency Cepstral Coefficients) parameters. Preliminary experiments showed that addition of derivatives (Δ) and accelerations ($\Delta\Delta$) of these features improved the performance. Thus, we use a 18-dimensional acoustic feature vector (see Table 1). Pitch, power and spectral features can be regarded as independent, thus are processed as three different streams in HMM. Preliminary experiments also showed that modeling the distribution with a mixture of eight Gaussians brought the best result.

F_0 is extracted using short time auto-correlation analysis of the residual signal by PARCOR analysis. In voiced segments, candidates of F_0 contours are narrowed down with beam search and continuity constraint. In unvoiced segments, 3-dimensional spline interpolation is applied based on the result of neighboring voiced segments. Logarithms

Table 1. Specification of HMM for sentence stress

acoustic features (18 dimension)	log(F_0), log(power), 4th-order MFCC their Δ and $\Delta\Delta$
analysis frame	frame period: 10ms, frame length: 25ms
HMM structure	left-to-right 3-state, 8-Gaussian mixture 3 stream (F_0, power, MFCC)

[1] V denotes Vowel and C denotes Consonant

of F_0 and power are computed every 10ms and normalized by the difference between the absolute maximum and the minimum in the whole contour.

3.2. Syllable Alignment by Phone Recognition

In order to reliably align the syllable sequence with handling phone insertions and substitutions by non-native speakers, we make use of a speech recognition system with error prediction for a given sentence. A list of segmental errors common among Japanese speakers of English was built for the prediction so that the speech recognizer can identify erroneous syllables and also pauses inadequately inserted between words[10].

3.3. Syllable Stress Recognition

Based on the alignment, the syllable units are constructed and their structures and positions in the phrase are determined. For each syllable, NS, PS and SS models are applied to determine the stress level. Linguistic studies suggest that all syllables but one in a word tend to be un-stressed in continuously spoken sentences[7]. Hence, we constrain the number of stressed (PS or SS) syllables to be one per word. Moreover, as defined in section 2.3, we constrain the number of PS to be one per phrase unit.

The most probable stress (syllable) sequence by the matching with the aligned phrase segment is obtained. It is compared with the correct stress pattern. Syllables whose detected stress level differs from the correct one are labeled as pronunciation errors. If the syllable structure and/or the position in the phrase are incorrect, such information is presented to the student as possible causes of the stress error.

3.4. Two-Stage Recognition of Syllable Stress

Recognition with HMM is based on Viterbi score $f(i,t)$ which is accumulatively computed for HMM state i and time t as:

$$f(i,t) = \max_j \left[f(j,t-1) \cdot a_{ji} \cdot b_i(y_t) \right]$$

where a_{ji} is a state transition probability from j to i and $b_i(y_t)$ is a probability density function of feature vector y_t at state i. As F_0, power and MFCC are modeled by different streams y_t^1, y_t^2, y_t^3 respectively, the emission probability $b_i(y_t)$ is re-written as:

$$b_i(y_t) = \{b_i^1(y_t^1)\}^{w_1} \cdot \{b_i^2(y_t^2)\}^{w_2} \cdot \{b_i^3(y_t^3)\}^{w_3}$$

where $b_i^x(y_t^x)$ is an emission probability and w_x is a weight of each stream. The weights are estimated by discriminant analysis using whole syllable segments as described in [11].

Since PS, SS and NS have different acoustic characteristics, the weights will be different according to the stress level. For example, PS is supposed to be characterized by a tonal change, thus F_0 should be the most important feature for discrimination. We propose a two-stage recognition method that applies different weights at each stage as shown in Figure 2. At the first stage, the presence of stress is detected. Here, a stress model (ST) that merges PS and SS syllables is compared against NS using weights optimized for the two-class discrimination. For syllables detected as stressed, the stress level (PS or SS) is recognized at the second stage using different weights.

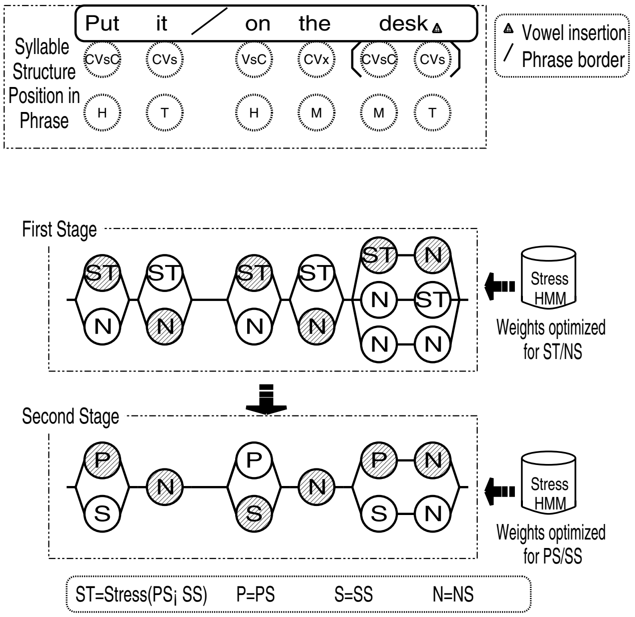

Fig. 2. Two-stage sentence stress recognition

Table 2. Breakdown of stress syllables in training and test data

speakers	syllable	PS	SS	NS
native speakers	3880	572 (15%)	1045 (27%)	2263 (58%)
Japanese students	935	208 (22%)	169 (18%)	558 (60%)

4. EXPERIMENTAL EVALUATION

We use a portion of TIMIT database (31 native speakers from a geographical region (dr1)) for training the models. Speech database of Japanese students used for evaluation consists of 108 sentences by 8 speakers (4 male, 4 female). We asked a native expert phonetician to label the stress level.

By comparing the distribution of stress levels by Japanese students and native speakers listed in Table 2, it is observed that the ratio of primary stress (PS) by Japanese students is far larger than that by native speakers though the percentage of non-stressed syllable (NS) is much the same. This is consistent with the observation that Japanese students

Table 3. Stress recognition accuracy with various models

category	#model	stream weight	native	Japanese
base	(3)	(1.0,1.0,1.0)	69.1%	59.3%
syl	(48)	(1.0,1.0,1.0)	80.0%	65.3%
pos	(9)	(1.0,1.0,1.0)	77.9%	61.8%
syl+pos	(144)	(1.0,1.0,1.0)	93.7%	79.3%

Table 4. Weights estimated by discriminant analysis (LDA)

target	F_0	power	MFCC	duration
ST vs NS	0.170	0.389	0.222	0.218
PS vs SS	0.520	0.175	0.179	0.126
PS vs SS vs NS	0.282	0.306	0.217	0.195

tend to overuse pitch in marking stress.

By another analysis of the distribution of stress levels, we also confirmed that (1) complex syllables are more likely to be stressed than simple syllables, as previously shown in [9], and (2) syllables at the end of phrases are more often stressed than others, which is also conventionally pointed out. The results support that the categories based on the syllable structure and position in a phrase are reasonable.

Baseline recognition results by setting all stream weights to 1.0 are listed in Table 3. The results for native speakers are computed with cross-validation, where three speakers were used as test set and the remaining 28 as training. Using detailed categories based on the syllable structure (syl) and position in a phrase (pos) leads to significant improvement of accuracy. The best results both for native (93.7%) and Japanese (79.3%) are obtained by combining all factors.

Then, we performed linear discriminant analysis (LDA) to compute the weights to discriminate between (1) ST-NS and PS-SS (two-stage) and (2) PS-SS-NS (one-stage), respectively. The estimated weights are listed in Table 4. For discrimination of ST and NS, power and vowel quality information (MFCC) are important, while F_0 is dominant for discrimination of PS and SS. The result confirms that PS is more characterized by phrasing information and pitch movement. With these weights, we identify the stress by two stages. For comparison, we also tried one-stage recognition using PS-SS-NS weights. Recognition results are shown in Table 5.

With the one-stage recognition, no improvement is observed in spite of using the weights estimated by LDA. On the other hand, the two-stage recognition method improves accuracy by 1.4% for native speakers and 4.8% for Japanese students. The result confirms effectiveness of the proposed two-stage method that models PS and SS in different ways.

Table 5. Recognition accuracy with various stream weights

model	method	weight	native	Japanese
syl+pos	one-stage	all 1.0	93.7%	79.3%
syl+pos	one-stage	LDA	93.5%	78.7%
syl+pos	two-stage	LDA	95.1%	84.1%

5. CONCLUSIONS

We have proposed a method to evaluate sentence stress in English spoken by Japanese students. Stress models are set up according to stress level, syllable structure and position of the syllable in a phrase. Three streams of acoustic features (pitch, power and MFCC) are used, and their weights are estimated by discriminant analysis. Decomposing stress recognition into two stages, one to detect the presence of stress and the other to determine the stress level, improved recognition accuracy together with optimization of the weights at each stage. The method achieved accuracy of 95.1% for native and 84.1% for non-native speakers.

6. REFERENCES

[1] S.Witt and S.Young. "Language Learning based on Non-Native Speech Recognition". In *Proc. ICASSP*, 1997.

[2] G.Kawai, A.Ishida, and K.Hirose. "Detecting and correcting mispronunciation in non-native pronunciation learning using a speech recognizer incorporating bilingual phone models". *J. Acoustical Society Japan*, 57(9):569–580, 2001.

[3] F.Ehsani and E.Knodt. "Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm". *Language Learning and Tecgnology*, 2(1):45–60, 1998.

[4] N.Minematsu, Y.Fujisawa, and S.Nakagawa. "Automatic detection of stressed syllables in English words using HMM and its Application to Prosodic Evaluation of Pronunciation Proficiency". *Trans. IEICE*, J82-D-II(11):1865–1876, 1999.

[5] G.J.Freij, F.Fallside, C.Hoequist, and F.Nolan. "Lexical Stress Estimation and Phonological Knowledge". In *Speech Communication and Language*, number 4, pages 1–15, 1990.

[6] K.L. Jenkin and M.S. Scordilis. "Development and Comparison of Three Syllable Stress Classifiers". In *Proc. ICSLP*, 1996.

[7] K. Watanabe. *"Instruction of English Rhythm and Intonation"*, chapter 4-6. Taishukanshoten, 1994.

[8] M. Sugito. *"English spoken by Japanese"*, chapter 1-4. Izumishoin, 1996.

[9] R.M. Dauer. "Stress-timing and Syllable-timing Reanalyzed". *Journal of Phonetics*, 11:51–62, 1983.

[10] Y.Tsubota, T.Kawahara, and M.Dantsuji. CALL system for Japanese students of English using formant structure estimation and pronunciation error prediction. In *InSTIL 2002 Advanced Workshop*, 2002.

[11] K.Imoto, M.Dantsuji, and T.Kawahara. "Modelling of the perception of English sentence stress for computer-assisted language learning". In *Proc. ICSLP*, 2000.