MODELLING OF THE PERCEPTION OF ENGLISH SENTENCE STRESS FOR COMPUTER-ASSISTED LANGUAGE LEARNING

Kazunori Imoto * *Masatake Dantsuji*† *Tatsuya Kawahara* *

* Graduate school of Informatics,
Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
† Center for Information and Multimedia Studies / Graduate school of Informatics,
Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

For learning foreign language pronunciation, prosodic features are important as much as, or more than segmental features. For Japanese speakers, one of difficulties to learn English pronunciation is rhythm because of the differences between two languages: mora-timing rhythm and stress-timing rhythm. In order to correct errors in rhythm, the method of evaluating sentence stress that constitutes rhythm is very significant. In this paper, we present a method of automatic detecting sentence stress syllables for the evaluation criterion. Using a linear discriminant function of pitch, intensity and vowel duration, about 90% of the syllables were correctly detected as to sentence stress. Also we analyzed the different and common characteristics among different English native speakers. The results revealed that the perception of the sentence stress among 11 native speakers had general agreement with respect to how to integrate three features.

1. INTRODUCTION

Prosodic features play an important role in human communications as to make clear focal point of topics, and also to emphasize or express one's intention. However, the differences between English and Japanese in the expressions of stress and rhythm cause serious difficulties for learners mastering prosodic patterns. Japanese speakers tend to utter English pronunciation in monotonous rhythm, which occasionally misleads their communication. So effective evaluation and instruction of rhythm are essential to acquire correct English pronunciation.

Recent advancement in speech technology enables us to develop a computer-assisted language learning (CALL) system. In the previous studies[1][2][3], effective evaluation criteria for segmental features or intonation were proposed. Also, the method of evaluating and instructing English word accent for Japanese was proposed[4]. However, the evaluation criterion of sentence stress, which is one of the most essential factors of English rhythm, is not established. So we propose a method of automatic detecting sentence stress syllables for a base of an effective CALL system.

It is known that English syllable stresses consist of pitch, intensity and duration. These features correspond to fundamental frequency, power and vowel duration, respectively. In [6], the rhythm instruction was realized by judging the stress with duration and vowel quality, not with pitch or intensity. According to [8], pitch and power also become key factors of sentence stress. So we utilized three features to detect sentence stresses. However, it is not clear which acoustic feature is the most important or how these features are integrated when English native speakers perceive sentence stresses. We investigate these issues and try to establish a universal evaluation criterion of English sentence stresses.

In order to estimate the appearance of sentence stress syllables from contours of fundamental frequency and power, each acoustic feature is to be normalized and quantized by syllable units. Then we adopt a linear discriminant function that integrates these acoustic features with weights. The weights, which reflect the importance of each acoustic feature in perceiving English sentence stresses, are estimated by discriminant analysis using the TIMIT database.

From the degree of agreement between natives' perception and our method by a linear discriminant function, we verify the validity of our model. Also by comparison with the weights that are estimated from different natives' labels, we see the common and different characteristics among different natives' perception.

2. SPEECH MATERIAL

As materials for the study, we picked up 310 sentences produced by English native speakers with New England dialect from the TIMIT database. Those speech samples were labeled by eleven English native speakers. From their perception, we eliminated inadequate speech samples that had neutral declarative rhythm to improve the reliability of speech materials and those labels.

Table 1: The number of syllables which each native evaluator perceives as sentence stress (Strong) or not (Weak).

Evaluator	A	B	C	D	E	F
Strong	310	548	397	473	384	578
Weak	1198	960	1111	1035	1124	930
Evaluator	G	H	I	J	K	Com
Strong	457	376	501	625	416	417
Weak	1051	1132	1007	883	1092	1091

As a result, we use 120 sentences that contain 1508 syllables for the experiment. Table 1 shows the result of the perception, which tells if each syllable is sentence stress (Strong) or not (Weak) by 11 native speakers. Out of them, 417 syllables, which were perceived by over 8 persons as sentence stress, were labeled as Strong and the remainder as Weak (Common label). Table 1 also suggests that there are differences among natives' perception. We verify this problem in the fourth section.

3. MODELLING OF ENGLISH SENTENCE STRESS

In this section, we describe our method of automatic detecting sentence stress syllables. Figure1 shows the procedure.

In the step1, we extract three prosodic features: fundamental frequency, power and vowel duration. Fundamental frequencies are extracted at 5 msec intervals using short-time autocorrelation analysis of the residual signal from the PARCOR analysis[7]. Additionally, the smoothing process and the linear interpolation in unvoiced segments are carried out for the robust extraction. In order to detect vowel duration and to quantize each feature by syllable units in the step2, we perform the phonetic alignment by the Viterbi algorithm.

Alternate appearance of sentence stress in syllable units constructs English rhythm. To detect sentence stress syllables from contours of fundamental frequency and power, these features are normalized and quantized by syllable units in the step2.

1. fundamental frequency
   The fundamental frequency contour consists of two elements: local peak and smooth curve. Because of this, the detection of sentence stress using the absolute pitch value is inadequate. In the previous study[8], it was shown that the pitch change correlated closely with sentence stress. Then we defined a new parameter " Degree of F_0 fall". Concretely, this is calculated as follows. At first, in each syllable, the difference between the local maximum and the next local minimum is calculated. (If there is no local maximum in the syllable, the value is to be zero.) Next, the difference is normalized by the difference between the absolute maximum and the absolute minimum in the whole F_0 contours. In order to confirm the correlation between Degree of F_0 fall and sentence stress, we conducted a preliminary experiment with speech materials. We compared several normalization method of fundamental frequency such as the average F_0 (Ave), the maximum F_0 (Max), the Z-averaged F_0 (Z-ave) and Degree of F_0 fall (Fall) in syllable units. Thresholds are determined so as to maximize each detection rate. Table 2 shows the results and suggests that the pitch fall is the most correlative.

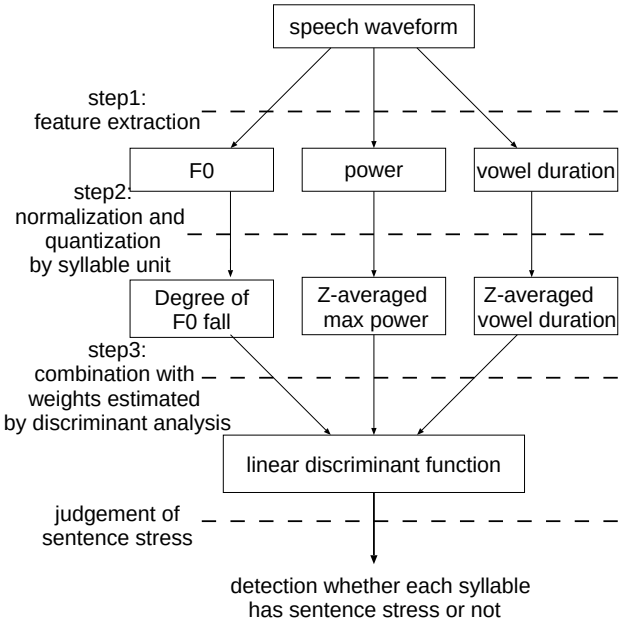

Figure 1: Process to detect sentence stress syllable

Table 2: Detection rate by using fundamental frequency

Parameter	Ave	Max	Z-ave	Fall
Correct	45.5(%)	61.8(%)	65.8(%)	82.3(%)

2. power
   The normalization of power is done as follows. At first, we calculate the average and the standard deviation value of power in the whole sentence. Then we compute the Z-averaged power with them and plot the maximum value in each syllable. The result of the preliminary experiment is shown in Table 3. This table shows that the Z-averaged power is more correlative than values normalized by the rest methods.

Table 3: Detection rate by using power

Parameter	Ave	Max	Z-ave	Fall
Correct	34.5(%)	45.3(%)	62.5(%)	53.8(%)

3. vowel duration
   The absolute vowel duration is affected by the speaking rate. To normalize the speaking rate, we calculate the difference between each vowel length and the average vowel duration. The average duration of each categorized vowel (short vowel, long vowel and diphthong) were calculated from the TIMIT database beforehand. In the preliminary experiment, we compared the absolute duration, Z-averaged duration (categorized into each vowel) and Z-averaged duration (categorized into short vowel, long vowel and diphthong). The result in Table 4 shows that the rate of the correct detection is improved by the normalization.

Table 4: Detection rate by using vowel duration

Parameter	Absolute duration	Z-ave (phone)	Z-ave (category)
correct rate	53.9(%)	65.2(%)	71.8(%)

Finally in the step3, three acoustic features are integrated as an evaluation criterion. We adopt a linear discriminant function. The weights used for the combination should reflect the importance in perceiving sentence stress by native speakers. For this reason, we analyze speech materials statistically. Concretely, we adopt the discriminant analysis to estimate the weight of each acoustic feature. By comparing the weights, how native speakers combine three features is judged. As a result, we can judge whether each syllable has sentence stress or not.

4. EXPERIMENT

4.1. Training and Evaluation by Common label

In order to verify the effectiveness of our method to detect sentence stress, we conducted two types of experiments. One is a closed test, in which we utilize all of the speech material both for training weights and evaluating sentence stress syllables. The other is an open test, in which 110 speech samples are used for training weights and the remainder for evaluation. In the open test, we conduct twelve experiments to replace speech samples for training and testing, and get their average. The evaluation results of test set by a linear discriminant function and the weights of the acoustic features are shown in Table 5. We compute the degree of agreement between natives' perception and our method using a linear discriminant function (Total Correct rate), and the Strong Correct rate, which is calculated by the geometric average between recall and precision of sentence stress syllables. Each weight in Table 5 is normalized so that the weight of F_0 is 1.000.

As a result of both experiments, the degree of agreement between natives' perception and our method is greater than 90%. Also, compared with the result of the preliminary experiments using only one acoustic feature in Table 2, the rate is improved by as much as 8%. These results show that the combination of three acoustic features using a linear discriminant function is a valid model of English sentence stress perception. According to these weights, the pitch change or fundamental frequency is the primary factor in the perception of sentence stress.

However, compared with the Weak Correct rate, the Strong Correct rate is lower by as much as 10%. English sentence stress may be classified into two categories: primary stress (PS) and secondary stress (SS). Primary stress can be assigned to most prominent syllables and secondary stress to the remaining stress syllables. In the previous study[5], the classification of two types of stresses (PS and SS) was tried. The result showed that the classification of SS was not easy. The two kinds of stresses are surmised to have slightly different characteristics. In this work, the difference is disregarded. But, even for English native speakers, it is hard to distinguish between two types of stresses.

4.2. Differences among natives' perception

The difference of the number of sentence stress syllables in Table 1 indicates that each native speaker may perceive sentence stress by using different criteria. In order to see the similarities and differences among different natives' perception, and to realize a universal evaluation criterion, we conducted further experiments. We labeled speech samples by different native evaluators and estimated the parameter weights individually. Each weight trained by different native perception reflects each speaker's tendency of sentence stress perception. The weights estimated by 11 native speakers are shown in Table 6.

From Table 6, we found general agreement among all native evaluators. The weights of F_0, power and vowel duration are similar. The values are about 1.00, 0.45 and 0.25, respectively. The only difference is in the threshold value. This shows that all native evaluators perceive sentence stress by the similar combination of the three acoustic features and our model realizes the universal evaluation criterion. The reason for the difference in threshold is that it is difficult to judge whether there is stress or not due to neutral declarative speech for some syllables.

Table 5: Detection rate by linear discriminant function and weight of the parameters

	Correct rate(%)			Weight (relative to F_0 weight)			
	Total	Strong	Weak	F_0	power	duration	threshold
closed	90.62%	81.35%	92.67%	1.000	0.438	0.269	0.433
open	90.12%	80.49%	92.25%	1.000	0.395	0.277	0.435

Table 6: Weights estimated by different natives' perception

	Weight (relative to F_0 weight)			
	F_0	power	duration	threshold
A	1.000	0.456	0.265	0.590
B	1.000	0.517	0.272	0.246
C	1.000	0.458	0.251	0.333
D	1.000	0.502	0.284	0.275
E	1.000	0.398	0.211	0.295
F	1.000	0.450	0.253	0.253
G	1.000	0.425	0.230	0.272
H	1.000	0.376	0.202	0.395
I	1.000	0.483	0.318	0.317
J	1.000	0.514	0.290	0.147
K	1.000	0.456	0.244	0.357

4.3. Evaluation of speech by Japanese speakers

Finally, we verify that our model is effective for English speech uttered by Japanese speakers. As speech materials, we collected 90 sentences that contain 822 syllables uttered by 12 male native Japanese speakers. The sentences were selected from the TIMIT database. Each syllable was judged by 11 English native evaluators to see if it was a sentence stress or not. We also evaluated each syllable by using the linear discriminant function. The weights used in this experiment were based on the result of the open test in Table 5. Table 7 shows that the degree of agreement is 83.1% in total. This result verifies that our model is effective for Japanese speakers, too. The reason for slight degradation in the Correct rate is due to pronunciation errors, which are typical for Japanese students with inserting vowels. Therefore we must take measures to reduce the effect caused by these errors when constructing the CALL system.

5. CONCLUSIONS

We have proposed a method of automatic detecting sentence stress syllables. We utilize three acoustic features of fundamental frequency, power and vowel duration, and combine them by a linear discriminant function. Each weight is trained by discriminant analysis.

The results show that, for speech samples uttered by both native speakers and Japanese, we could detect accurately 90% and 83% of the syllables, respectively. Also we confirmed general agreement among the perception of 11 native speakers with respect to how to integrate three features. These results show that our model can be a foundation of an effective CALL system with proper instruction.

Table 7: Evaluation for speech uttered by Japanese speakers

Correct rate(%)		
Total	Strong	Weak
83.1(%)	77.3(%)	87.4(%)

Acknowledgment: This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas, No.12040106, 2000.

References

[1] S. Auberg, N. Correa, V. Locktionova, R. Molitor and M. Rothenberg: The Accent Coach: An English Pronunciation Training System for Japanese Speakers In *ESCA - STiLL98*, pp.103-106, 1998.

[2] L. Neumeyer, H. Franco, M. Weintraub and P. Price: Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech In *Proc. ICSLP*, 1996.

[3] C. Jo, T. Kawahara, S. Doshita and M. Dantsuji: Automatic Pronunciation Error Detection and Guidance for Foreign Language Learning In *Proc. ICSLP*, 1998.

[4] Y. Fujisawa, N. Minematsu and S. Nakagawa: Evaluation of Japanese Manners of Generating Word Accent of English Based on a Stressed Syllable Detection Technique In *Proc. ICSLP*, 1998.

[5] K. Jenkin, M. Scordilis: Development and Comparison of Three Syllable Stress Classifiers In *Proc. ICSLP*, 1996.

[6] S. Hiller, E. Rooney, J. Laver and M. Jack: An Automated System for Computer-Aided Pronunciation Teaching In *Speech Communication*, 1993.

[7] H. Fujisaki, K. Hirose and S. Seto: A Scheme for Pitch Extraction of Speech Using Autocorrelation Function with Frame Length Proportional to Time Lag, In *Technical Report of IEICE*, SP90-86, pp.9-16 (1990, in Japanese).

[8] M. Sugito: English spoken by Japanese, published by Izumi shoin (1996, in Japanese)