Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 29–42

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Automatic sentence stress feedback for non-native English learners

Gary Geunbae Lee [a], Ho-Young Lee [b,*], Jieun Song [c], Byeongchang Kim [d], Sechun Kang [e], Jinsik Lee [f], Hyosung Hwang [b]

[a] *Department of Computer Science and Engineering, Pohang University of Science and Technology, 77 Cheongam-ro. Nam-gu, Pohang, Gyeongbuk, South Korea*

[b] *Department of Linguistics, Seoul National University, Daehak-dong, Gwanak-gu, Seoul, South Korea*

[c] *Speech Hearing and Phonetic Sciences, University College London, 2 Wakefield St., London WC1N 1PF, UK*

[d] *School of Computer and Information Communication Engineering, Catholic University of Daegu, Hayang-ro, Gyeongsan, Gyeongbuk, South Korea*

[e] *Software R&D Center, Samsung Electronics, Maetan-dong, Yeongtong-gu, Suwon, Gyeonggi, South Korea*

[f] *DMC R&D Center, Samsung Electronics, Maetan-dong, Yeongtong-gu, Suwon, Gyeonggi, South Korea*

Received 19 March 2014; received in revised form 11 April 2016; accepted 13 April 2016
Available online 6 June 2016

Abstract

This paper proposes a sentence stress feedback system in which sentence stress prediction, detection, and feedback provision models are combined. This system provides non-native learners with feedback on sentence stress errors so that they can improve their English rhythm and fluency in a self-study setting. The sentence stress feedback system was devised to predict and detect the sentence stress of any practice sentence. The accuracy of the prediction and detection models was 96.6% and 84.1%, respectively. The stress feedback provision model offers positive or negative stress feedback for each spoken word by comparing the probability of the predicted stress pattern with that of the detected stress pattern. In an experiment that evaluated the educational effect of the proposed system incorporated in our CALL system, significant improvements in accentedness and rhythm were seen with the students who trained with our system but not with those in the control group.

Keywords: Sentence stress; Sentence stress feedback system; Stress prediction model; Stress detection model; Stress feedback provision model; CALL

1. Introduction

In English, every word has one or more lexical stresses[1] depending on the structure of the word and the number of syllables, but not all word stresses are phonetically realized in utterance. Content words, which deliver major semantic information and therefore require listeners' attention, normally receive stress on the utterance level whereas function words do not (cf. Kingdon, 1958, for a detailed discussion about content and function words). Stress imposed on the utterance level has been traditionally called 'sentence stress' (Gimson, 1980; Jones, 1972).

* Corresponding author at: Department of Linguistics, Seoul National University, Daehak-dong, Gwanak-gu, Seoul, South Korea. Tel: +82 10 2744 0584; fax: +82 2 882 2451.

E-mail address: hylee@snu.ac.kr (H.Y. Lee).

[1] Each word has one primary stress.

http://dx.doi.org/10.1016/j.csl.2016.04.003

The major function of sentence stress is to highlight semantically important words and to form the rhythmic pattern of the utterance. It has been known that sentence stress occurs at regular intervals and unstressed syllables between consecutive stressed syllables are reduced, causing the impression of 'stress-timed rhythm'. Stress-timed rhythm has been traditionally distinguished from 'syllable-timed rhythm' in which syllables are pronounced with similar duration without vowel reduction, as in French and Italian (Abercrombie, 1967; Lloyd James, 1940; Pike, 1945). However, the idea that the units of rhythm (i.e., feet in stress-timed rhythm and syllables in syllable-timed rhythm) are of equal duration has been negated in several empirical studies (e.g., Dauer, 1983; Roach, 1982; Wenk and Wiolland, 1982) and for many languages, the rhythmic classification is not as clear-cut as previously believed. Nonetheless, the rhythm classes are generally regarded as reflecting the different rhythmic characteristics.

Sentence stress is distinguished from pitch accent that carries pitch prominence caused by an important intonation event (Bolinger, 1958; Pierrehumbert, 1980) as well as rhythmic prominence caused by sentence stress. While all pitch accents are imposed on stressed syllables, some sentence stresses do not involve pitch prominence and only affect the rhythmic pattern of a sentence. Thus pitch accent can be considered to be ranked higher than sentence stress in the prosodic hierarchy.

It has been recognized in some previous research that prosody plays an equal or greater role than segments in the judgment of comprehensibility and/or accentedness of non-native speech (e.g. Anderson-Hsieh et al., 1992; Boula de Mareüil and Vieru-Dimulescu, 2006; Hahn, 2004; Tajima et al., 1997; Wennerstrom, 2000). As for Korean learners of English, great difficulties have been observed in rhythm and fluency. Low proficiency learners tend to place sentence stress on most of the words in a sentence, even on function words (Lee, 2011). They tend to use strong vowels even in unstressed syllables, giving the impression of syllable-timed rhythm to native English listeners.

Previous research shows that teaching only the pronunciation of segments does not significantly improve comprehensibility in non-native spontaneous speech (Elliott, 1997), whereas prosody teaching does (Derwing and Rossiter, 2003; Derwing et al., 1998). Hence this paper aims to propose an automatic sentence stress feedback system designed to provide non-native English learners, especially Koreans, with feedback on their sentence stress errors. It is hoped that this system will help non-native learners correct their errors and thereby improve their English rhythm and fluency, ultimately resulting in an increase in their oral proficiency.

Automatic prosody scoring systems have been proposed in previous literature to evaluate the English prosody (i.e., stress, rhythm and/or intonation) of both native and non-native speakers (Cheng, 2011; Hönig et al., 2010; Ito et al., 2009; Liscombe, 2007; Maier et al., 2009; Mostow and Duong, 2009; Suzuki et al., 2008; Tepperman et al., 2010; Yamashita et al., 2005). Several automated speech assessment systems with a prosody evaluation component have also been developed (Chandel et al., 2007; Chen and Zechner, 2011; Teixeira et al., 2000; Zechner et al., 2011). These automatic prosody or speech assessment systems are useful for stimulating and encouraging English learners, but their educational effect is somewhat limited because corrective feedback is not offered and learners are left without knowing what to correct. Hence, there have been a few studies concerning systems that offer automated feedback on prosody or speech (Bang et al., 2013; Sitaram et al., 2011).

Imoto et al. (2002) proposed a sentence stress detection model that provides diagnostic information to learners by comparing native speakers' reference speech with Japanese learners' speech. This study focused only on stress detection. Our proposed system, however, focuses on the integration of sentence stress prediction, detection, and error feedback technologies to help non-native learners effectively improve their English rhythm and fluency.

While several previous studies including Imoto et al. (2002) used native speech as a reference for direct comparison with non-native speech, this study uses automatically predicted sentence stresses generated by our sentence stress prediction model as a reference for sentence stress detection. This allows the proposed system to evaluate and process any given utterance.

Frequent and repetitive practice is necessary for non-native learners to improve their English rhythm and fluency. Since traditional face-to-face language learning opportunities are costly due to time and space barriers, CALL (Computer-Assisted Language Learning), which overcomes these barriers and offers non-native learners easy access to computer-based practice programs, has received much attention since Levy (1997) and Witt and Young (1997). Lee et al. (2011) showed that a CALL system based on dialog management technologies (cf. Lee et al., 2010) significantly improved learners' spoken language skills. To evaluate our sentence stress feedback system, we set up a CALL system with which our sentence stress prediction, detection, and feedback provision models were incorporated.

The remainder of this paper is structured as follows: Section 2 describes the materials used for the proposed system. Section 3 proposes and describes the sentence stress prediction, detection, and feedback provision models. Section 4

Table 1
A comparison of Foot and Narrow Rhythm Unit.

Sentence	It’s	ˈalmost	im-	ˈpossible
Abercrombie	[$_{\text{F}}$ It’s]	[$_{\text{F}}$ almost im-]		[$_{\text{F}}$ -possible]
Jassem	[$_{\text{ANA}}$ It’s]	[$_{\text{NRU}}$ almost]	[$_{\text{ANA}}$ im-]	[$_{\text{NRU}}$ -possible]

reports the results of some validation experiments designed to evaluate the performance of these models as well as the usefulness of our CALL system in a real learning environment. A conclusion is given in Section 5.

2. Materials

2.1. The Aix-MARSEC database for sentence stress prediction

In Kang et al. (2012), we used the BURNC (Boston University Radio News Corpus), where pitch accents are labeled according to the ToBI system (Silverman et al., 1992) to produce a pitch accent prediction model. The accuracy of sentence stress prediction with this model was only 83.7%, and thus not satisfactory enough to be used for teaching stress-timed English rhythm to non-native learners because pitch accent is imposed on some, but not all, stressed words in a sentence.[2] Since we wanted to drastically improve the accuracy of the sentence stress prediction, we built a sentence stress prediction model using the Aix-MARSEC (Aix-Machine Readable Spoken English Corpus) database where sentence stress is reliably annotated (cf. Hirst et al., 2009).

The Aix-MARSEC database consists of over five hours of natural-sounding British English speech data collected from 53 different speakers (17 males and 36 females). The corpus includes approximately 55,000 orthographically transcribed words. In this corpus, rhythm units and stress feet, both of which are demarcated by sentence stress, are annotated.

The rhythm unit annotation is based on Jassem’s (1952) notion of Anacrusis (ANA) and Narrow Rhythm Unit (NRU) and the stress foot annotation on Abercrombie’s (1964) notion of Foot (F) (cf. the *Read Me* file available in Hirst et al., 2009). Narrow Rhythm Unit begins with a stressed syllable and ends at the following word boundary.[3] Foot begins with a stressed syllable or an intonation boundary and ends before the following stressed syllable or at the next intonation boundary. Hence the right boundary of the Narrow Rhythm Unit coincides with the word boundary whereas that of the Foot is often placed inside a word before the next stressed syllable. Unstressed syllables or words preceding the first stressed syllable are regarded as Anacruses, which are “pronounced extremely rapidly” according to Jassem (1952). But in Abercrombie (1964), utterance-initial unstressed syllables or words preceding the first stressed syllable form a separate Foot. The difference between the two theories of rhythm is well demonstrated in Hirst et al. (2009) as in Table 1.

If we regard the first syllable in NRU as a stressed syllable, we can easily extract words where sentence stress is imposed. But if we use Abercrombie’s stress feet annotation, we have to exclude utterance-initial feet without a stressed syllable, which requires extra effort. Hence we used Jassem’s rhythm unit annotations to implement our sentence stress prediction model. By using the Aix-MARSEC database, we were able to acquire a much more accurate sentence stress prediction model, which works for any sentence.

2.2. The KLEAC database for sentence stress detection

Non-native English learners have their own prosodic habits and characteristics depending on their language background. To increase the accuracy of sentence stress detection from learners’ speech, we needed to adapt the prosodic characteristics of non-native learners’ speech to our sentence stress detection model. In this model, we focused on sentence stresses imposed by Korean learners when they read English sentences. We used the KLEAC (Korean

[2] Although the pitch accent prediction model is not suitable for teaching English rhythm, it may be used to build an intonation assessment and correction system.

[3] It should be noted that Jassem (1999) defined ‘stress’ as ‘the potential for accent’ and used the term ‘accent’ to refer to both rhythm accent (tertiary accent) and pitch accent (primary and secondary pitch accents). ‘Accent’ in Jassem (1999) corresponds to ‘stress’ in Jassem (1952).

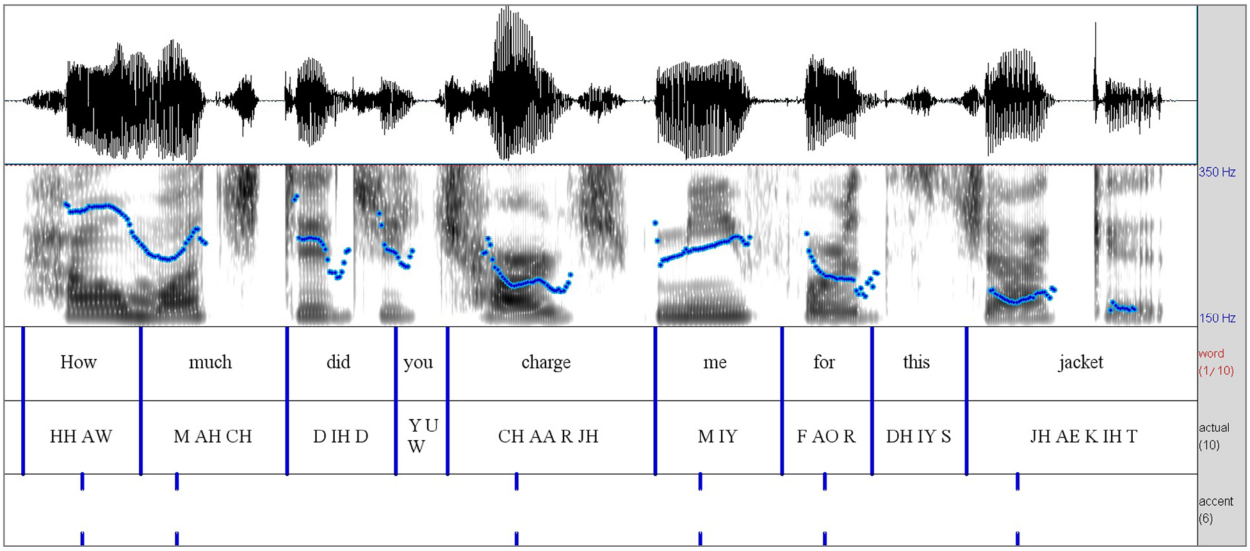


Fig. 1. A sample from the KLEAC.

Learners' English Accentuation Corpus) database, where sentence stresses imposed by Korean learners are annotated (cf. Lee, 2011) as displayed in Fig. 1. Naturally, some sentence stresses were found to be incorrectly placed, especially on function words.

The KLEAC consists of six hours of speech with 5500 English sentences produced by 75 native Korean speakers (middle school students aged 13–14 years). In this corpus, sentence stress labels, but not RNU/ANA labels, were manually annotated by seven Korean phonetic experts using the Praat program (Boersma and Weenink, 2009). A set of annotation principles were established; most importantly, sentence stress was marked on syllables produced with pitch prominence, longer duration, and unreduced vowel quality. The labelers were instructed to pay greater attention to function words because imposing sentence stress on them is one of the major errors made by Korean learners of English. To improve the inter-transcriber agreement, the annotations were partly cross-checked between the labelers in the beginning stage. The inter-rater agreement was calculated for 524 sentences that were randomly selected from the corpus using Fless' Kappa; there was very strong inter-rater agreement among the labelers [$\kappa = .868$ (95% CI: .866 to .870), $p < .001$].

To verify the reliability of the KLEAC labelers, five of them took part in an additional labeling task;[4] they were asked to mark sentence stresses imposed by English native speakers (not by Korean learners) on 50 utterances arbitrarily selected from the Aix-MARSEC database. Inter-rater agreements were then calculated between the Korean transcribers and the annotator(s) of the Aix-MARSEC database. Specifically, sentence stress labels made by each Korean labeler were compared with those made by the transcriber(s) who annotated the 50 chosen Aix-MARSEC utterances. The results of Cohen's Kappa analyses performed between each of the labelers and the Aix-MARSEC labeler(s) are shown in Table 2. Kappa values ranged between 0.61 and 0.83, which suggests that the agreement between the Korean labelers and the annotator(s) of the Aix-MARSEC database was substantial. Furthermore, a Fless' Kappa analysis demonstrated consistency among the Korean labelers [$\kappa = .721$ (95% CI: .713 to .729), $p < .001$]. The results of this labeling task indicate that the Korean labelers also transcribe native English sentence stress in a consistent manner.

3. Methods

3.1. Overview of the sentence stress feedback system

As mentioned above, our sentence stress feedback system aims to provide non-native learners with corrective feedback on English stress errors. This will allow learners to check and correct their own stress errors, which would

[4] Two labelers could not participate in this task because of personal reasons.

Table 2
The agreement rates between individual KLEAC labelers and Aix-MARSEC labelers.

Transcribers	1	2	3	4	5
Cohen's Kappa	κ = .702 95% CI: .631 to .774 p < .001	κ = .789 95% CI: .728 to .851 p < .001	κ = .617 95% CI: .539 to .695 p < .001	κ = .830 95% CI: .773 to .886 p < .001	κ = .829 95% CI: .772 to .885 p < .001

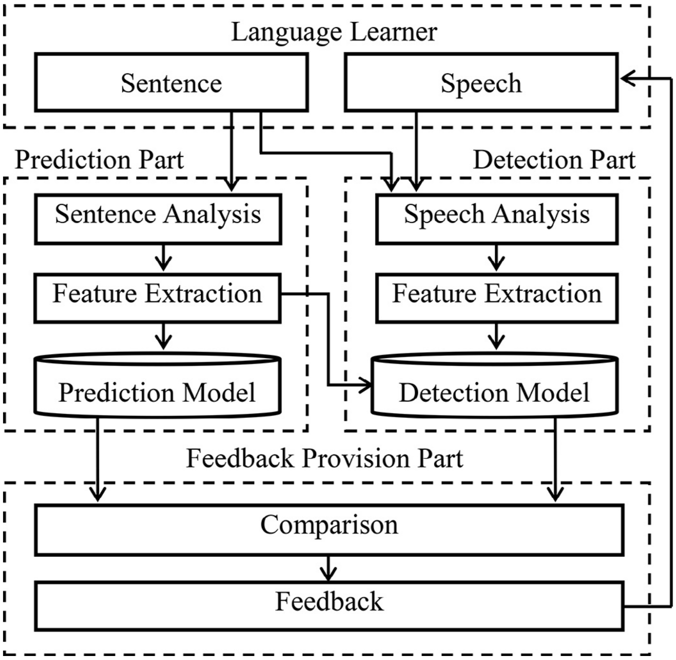


Fig. 2. Architecture of the proposed system.

eventually lead to the improvement of their English rhythm and fluency. To achieve this goal, we built an integrated sentence stress feedback system consisting of sentence stress prediction, detection, and feedback provision models, as in Fig. 2. The prediction and detection models were established using two different speech databases: the Aix-MARSEC database was used for the prediction model and the KLEAC for the detection model. In addition we developed a CRF (Conditional Random Field) classifier based on Lafferty et al. (2001).

If a practice sentence is given to the system, the prediction model produces a reference sentence stress pattern that is generated on the basis of some syntactic and lexical features. These features will be discussed in Section 3.2. When a non-native learner reads the practice sentence, the detection model analyses, extracts, and utilizes acoustic features to detect sentence stresses imposed on the utterance. By comparing the predicted sentence stresses with the detected ones, the feedback provision model informs the learner whether each word is stressed or unstressed correctly. In the proposed system, we carefully set up a feedback provision model so that it offers positive or negative stress feedback for every spoken word by comparing the probability of the predicted sentence stress pattern with that of the detected stress pattern.

The proposed system is designed to detect sentence stress errors on the word level, but not on the syllable level, because sentence stress detection processed on the syllable level considerably lowers the accuracy of the system. Hence this system cannot detect lexical stress errors.

The rates of sentence and lexical stress errors were measured as part of another study using 525 sentences arbitrarily chosen from the KLEAC corpus. The rate of lexical stress errors was only 0.4% (21 words out of total 4795 words) whereas the rate of sentence stress errors of function words was 20.7% (483 words out of 2333 function words). It follows that the inability of our system to detect lexical stress errors does not pose a serious problem in teaching English rhythm to Korean learners.

Although the proposed system is designed to detect sentence stress errors of both content and function words, sentence stress errors of content words are rarely detected unless a content word is pronounced very weakly. This system is also designed to detect sentence stress errors in utterances of broad focus. Hence it is very likely that our system will judge sentence stress placed on a narrowly focused function word as an error.

3.2. Sentence stress prediction model

To construct a stress prediction model that automatically generates a reference stress pattern for any input sentence, we used the Aix-MARSEC database. The first syllable of each NRU in this database was regarded as having attracted sentence stress on the utterance level.

In building our sentence stress feedback system for teaching English rhythm and fluency, we wanted the prediction model to accurately predict sentence stress placement when any new sentence was given without additional information. We also noted that the prediction of a word's sentence stress label was determined by the labels of neighboring words, not those of distant words. Hence we adopted the linear-chain CRF model which has been widely used in the previous studies on the automatic prediction of pitch accent, sentence stress, and boundary tones (Jeon and Liu, 2009; Qian et al., 2010; Rangarajan Sridhar et al., 2008).

The linear-chain CRF defines a conditional probability distribution of a label sequence *y*, given an observation sequence *x*. The distribution follows the relation between the labels encoded in a linear-chain structure. A linear-chain CRF is defined as follows:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k (y_{t-1}, y_t) + \mu_k g_k (y_t, x) \right) \tag{1}$$

where $\mathbf{Z}(\boldsymbol{x})$ is a normalization constant that is computed by summing over all possible label sequences $\boldsymbol{y}$ of the observation sequence $\boldsymbol{x}$. $f_k(y_{t-1}, y_t)$ encodes a transition score for state transition $y_{t-1} \rightarrow y_t$ and $g_k(y_t, \boldsymbol{x})$ encodes the observation feature, $\boldsymbol{x}$, centered at the current time, t. λ_k and μ_k are trained parameters associated with the features f_k and g_k for a given feature k.

Since content words tend to be stressed on the sentence level whereas function words are not, the word class, i.e. content or function word, was considered to be the most important feature among the several syntactic and lexical features that were generated based on POS (Part-of-Speech) tags automatically obtained by using our own POS tagger. This tagger was implemented following Brill (1992) and showed an accuracy of 96.3%.

Based on the POS tags, we generated new syntactic and lexical features by combining them with a syntactic or lexical feature like word class, word identity, number of vowels, and number of syllables. These new features are WORD_CLASS+POS_TAG, WORD_IDENTITY+POS_TAG, VOWEL_ NUMBER+POS_TAG, and SYLLABLE_NUMBER+POS_TAG.

The linear-chain CRF needs contextual information to appropriately predict the labels (stressed or unstressed) of individual words in a practice sentence. The two words preceding a target word and the three words following it were chosen as its contextual information. The optimal window size for this contextual information was empirically determined in the experiment.

3.3. Sentence stress detection model

Like the sentence stress prediction model, the sentence stress detection model was constructed with the same CRF classifier. While the prediction model was trained with the Aix-MARSEC database, we used the KLEAC database to train the detection model. This was done to reflect the acoustically unique characteristics of Korean learners' English so that the detection model could detect sentence stress in Korean English more accurately.

Much like the classification task of the prediction model, an important part in developing the detection model was determining the necessary features. Acoustic features were essential in detecting imposed sentence stresses from learners' utterances. The acoustic features used for stress detection included syllable duration, vowel duration, the normalized mean pitch of the syllable, the normalized mean intensity of the syllable, and the duration of silence at the word boundary.

According to previous research into prosody detection (Jeon and Liu, 2009; Qian et al., 2010; Rangarajan Sridhar et al., 2008), appropriate syntactic and lexical features are also necessary to achieve high accuracy of prosody detection. Hence the same syntactic and lexical features used for the prediction model were used to develop the detection model.

In the proposed stress detection model, the normalized pitch and intensity means were extracted from the mean pitch and intensity of lexically stressed vowels because sentence stress is normally imposed on syllables receiving primary word stress. To determine the timeline of the words in a given utterance, a forced alignment procedure was used.

However, non-native English has various acoustic characteristics that prevent the stress detection model from detecting sentence stress accurately. To reduce undesirable fluctuations, acoustic features were normalized using the *z*-score. The *z*-score was calculated using the mean μ and the standard deviation σ of the feature value *x*, as shown in the following sentence: $z = (x - \mu) / \sigma$.

Generally, the acoustic features were measured with continuous values. Before applying the continuous values to the CRF classifier, we discretized the continuous feature variables. To conduct quantile discretization, we introduced 10 bins where each bin received an equal number of data values. Through the quantile discretization procedure, we were able to alleviate the data sparseness problem.

3.4. Sentence stress feedback provision model

The probability scores of a word's predicted and detected stress levels were produced by the CRF classifier used in the stress prediction and detection models. These scores, which range from 0 to 1, reflect how confidently the models assess the stress level of each word. Sentence stress feedback signs are determined by comparing the stress probability scores produced by the prediction and detection models, word by word. As the absolute difference between the predicted and detected probability values of a word decreases, the detected stress level is assumed to correspond better to the predicted one. For example, for a word in a practice sentence, if the output probability score of the stress prediction model is 0.8 (likely to be stressed) and that of the stress detection model 0.2 (unlikely to be stressed), then the absolute difference is 0.6 (=|0.8 − 0.2|) and therefore the learner's stress placement on the word is likely to be different from the predicted standard stress pattern.

Two feedback signs were used: 'O' was a positive feedback sign, meaning 'right', and 'X' was a negative feedback sign, meaning 'wrong', as shown in Fig. 3. Positive feedback was used to encourage the learners' good performance. Negative feedback was used to inform the learners of their stress errors that need correction. Since both prediction and detection models can make errors in automatically judging word stress level, we had to consider the possibility of incorrect feedback signs being given to the learners. To avoid this problem and achieve reliability in the proposed system, the feedback model was designed so that no signs are offered if the absolute difference between predicted and detected probability scores falls within the range where neither positive nor negative feedback can be given confidently to the learners.

The formula for the feedback is as follows:

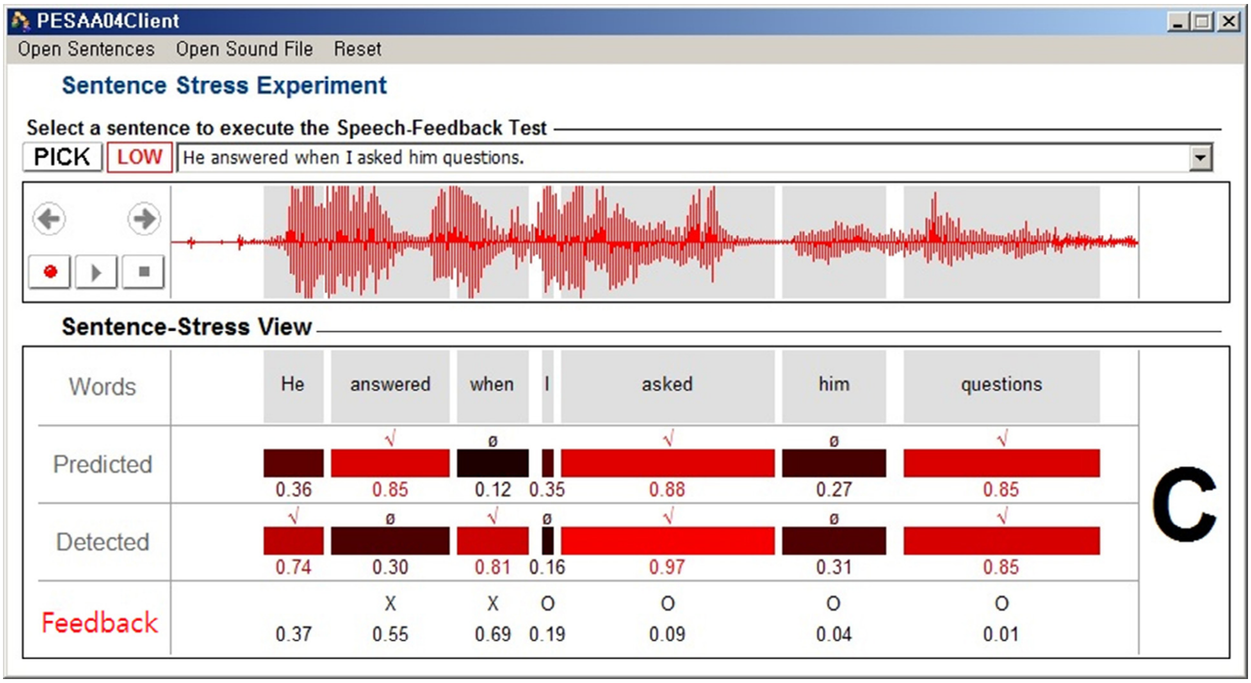
$$\text{Feedback} = \begin{cases} \text{Positive}, & \text{if } |\pi_{pre} - \pi_{det}| < \theta_1 \\ \text{No sign}, & \text{if } \theta_1 \le |\pi_{pre} - \pi_{det}| \le \theta_2 \\ \text{Negative}, & \text{if } |\pi_{pre} - \pi_{det}| > \theta_2 \end{cases} \tag{2}$$

The π_{pre} and π_{det} in (2) are output probability values derived by the prediction and detection models, respectively, and θ_1 and θ_2 are decision boundaries for the feedback signs.

4. Results

4.1. Evaluation of the sentence stress prediction and detection models

To validate the performance of the stress prediction and detection models, the utterances in the Aix-MARSEC and KLEAC databases were divided either into a test or a training set, respectively, in a ratio of approximately 4:1. Using



PESAA04Client

Open Sentences Open Sound File Reset

Sentence Stress Experiment

Select a sentence to execute the Speech-Feedback Test

PICK LOW He answered when I asked him questions.

Sentence-Stress View

Words	He	answered	when	I	asked	him	questions
Predicted	0.36	√ 0.85	ø 0.12	0.35	√ 0.88	ø 0.27	√ 0.85
Detected	√ 0.74	ø 0.30	√ 0.81	ø 0.16	√ 0.97	ø 0.31	√ 0.85
Feedback	0.37	X 0.55	X 0.69	O 0.19	O 0.09	O 0.04	O 0.01

C

Fig. 3. A sample of sentence stress feedback.

Table 3
Accuracy, precision, recall and F-measure of the proposed models.

Models	Accuracy	Precision	Recall	F-measure
Prediction	96.6	98.3	96.1	97.2
Detection	84.1	84.8	88.8	86.7

the datasets, we conducted five-fold cross validation for the prediction and detection models to calculate the accuracy, precision, recall, and F-measure values of these models, as in Table 3, and verified how accurately the proposed models predict and detect sentence stress. These values had been checked first because the proposed system manipulated the binary classification problem in which moderate accuracy could be achieved by selecting either stressed or unstressed words as holding a majority. Since the Aix-MARSEC database used for the prediction model showed an accuracy of 60.5% when selecting stressed words as holding a majority and the KLEAC database used for the detection model was 58.6% accurate under the same condition, we selected stressed words as holding a majority. To improve the accuracy of the prediction and detection models from the baseline accuracy, we adopted the machine learning method using a linear-chain CRF. As a result, the accuracy of both models was greatly improved as can be seen in Table 3.

Compared to our previous pitch accent prediction model (Kang et al., 2012), with which the accuracy of sentence stress prediction was only 83.7% [5], sentence stress prediction accuracy greatly improved because the Aix-MARSEC database offered much more appropriate information about sentence stress placement. Although the accuracy of the detection model was much lower than that of the prediction model, the accuracy and F-measure of the detection model reached 84.1% and 86.7%, respectively. It led us to believe that the detection model could be effectively used in a real teaching environment if the decision boundaries were appropriately assigned in the feedback provision model.

[5] The accuracy and F-measure of pitch accent prediction were respectively 87.3% and 88.3%, based on the BURNC.

Table 4
Comparison of labeled, predicted and detected results.

a. Prediction

	Predicted	
	Unstressed	Stressed
Labeled		
Unstressed	3670	96
Stressed	224	5521

b. Detection

	Detected	
	Unstressed	Stressed
Labeled		
Unstressed	3047	889
Stressed	626	4959

Table 4 shows the confusion matrices of the prediction and detection results. It shows the number of words 'labeled' as 'stressed' or 'unstressed' in the KLEAC and the number of words 'predicted' or 'detected' as 'stressed' or 'unstressed' with our prediction and detection models.

We can confirm from Table 4 that the prediction model results in higher accuracy than the detection model and that the error rate of stressed word detection is reasonably low, i.e. 11.2% (= 626 / (626 + 4959)). It follows that reliable sentence stress prediction and detection results are attainable with our prediction and detection models.

4.2. Determining feedback decision boundaries

Unlike the prediction and detection models, a validation of the feedback provision model should be based on how optimally the decision boundaries for feedback are decided so that it can help non-native learners correct sentence stress errors. Iterative calculations to find optimum boundaries were conducted until the correlation between the learners' feedback scores and their proficiency levels were maximized. If the feedback given to a learner was highly correlated with this learner' proficiency level based on the corresponding feedback decision boundaries, we verified that the feedback provision model with these boundaries were reliable and hence useful in a real learning environment.

To measure the appropriateness of the feedback provision model, we calculated Pearson's correlation coefficients between feedback-derived scores (Eq. 3) and the proficiency levels of the Korean learners that participated in the KLEAC database recordings.[6]

$$Feedback\ driven\ score = \frac{Number\ of\ "positive"\ \ feedback}{Total\ number\ of\ "positive"\ and\ "negative"\ feedback} \tag{3}$$

The utterances in the KLEAC database were split into a ratio of 4:1. Four fifths of all utterances were used to construct a new detection model and this model was then applied to the remaining one fifth of the utterances to derive the feedback-derived scores.

To measure the correlation between feedback scores and the proficiency levels of the Korean learners recorded in the KLEAC database, a five-fold cross validation process was conducted cyclically 5 times to avoid overfitting. Through the 5 iterative processes, each utilizing a new detection model, the feedback scores eventually took all of the KLEAC utterances into consideration; these scores were then correlated to the speakers' proficiency levels.

An iterative search for sentence stress was performed to find the θ_1 and θ_2 values that yielded the highest correlation between feedback scores and speaker proficiency levels. This was done by changing the θ_1 and θ_2 values in increments of 0.01, between 0 and 1. As a result, the search found that the highest correlation value of 0.5855 was achieved when θ_1 and θ_2 were respectively set as 0.18 and 0.61, as can be seen in Fig. 4.

[6] The proficiency levels were assessed by 7 native speakers. They were graduate students majoring in linguistics.

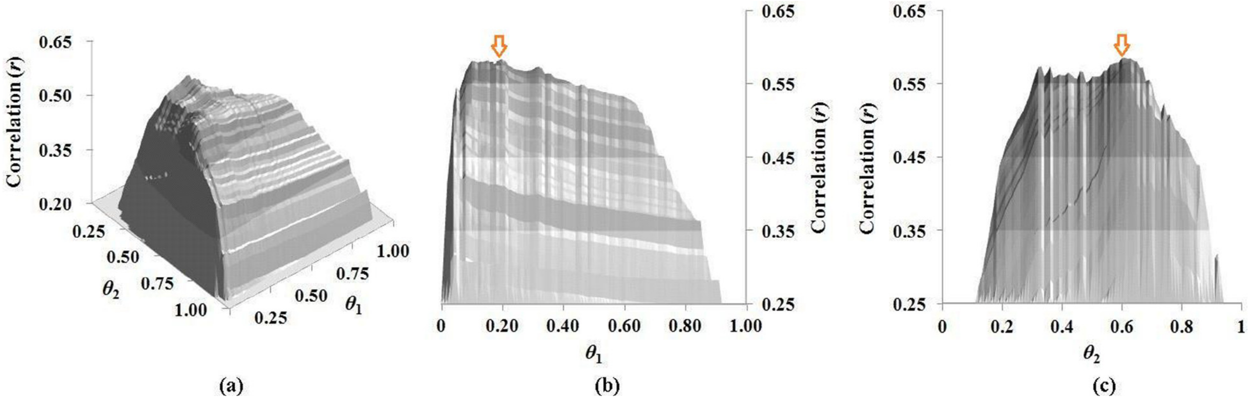


Fig. 4. Correlations between the learners' proficiency levels and the feedback scores determined by θ_1 and θ_2; (a) for an overview, (b) for θ_1 and (c) for $\theta_{2.}$

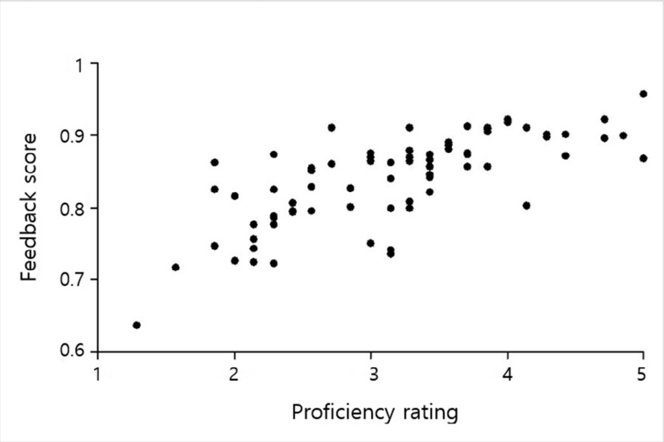


Fig. 5. Distribution of feedback scores over proficiency levels.

Fig. 5 displays the distribution of the feedback scores over the proficiency levels, based on the optimum boundaries. This figure shows that the feedback scores increase with the proficiency levels. This leads us to the assertion that our stress feedback provision model offers reasonably good sentence stress feedback, with which non-native learners can correct their stress errors and thereby improve their English rhythm and fluency.

4.3. Educational effect of the proposed system

To evaluate the educational effect of the proposed sentence stress feedback system, we conducted an additional training experiment by building an online CALL system in which our sentence stress feedback system was embedded. 40 Korean elementary school students consisting of 19 boys and 21 girls participated in this experiment. 38 of the students were approximately 12 years old (6th graders) and 2 of them were 11 years old (5th graders). They were enrolled in an English course at a private English institute in Seoul, Korea. Most of the students were assessed by the institute to be between lower and upper intermediate proficiency levels and their experience living in an English speaking environment was limited; only two students had experience residing in an English speaking community and their length of residence was less than six months.

During the training period, all 40 students took English classes both at school and at the institute. As a part of their homework for the institute course, the students were requested to solve listening problems and repeat after native speakers' read sentences. These exercises were completed online. In addition, they received High Variability Phonetic Training.

Table 5
Results of the assessment (mean and standard deviation).

	Comprehensibility		Accentedness		Rhythm	
	Before	After	Before	After	Before	After
Control group	5.47 (1.60)	5.72 (1.55)	4.25 (1.71)	4.45 (1.61)	4.15 (1.73)	4.38 (1.86)
Experiment group	5.55 (1.78)	5.75 (1.80)	4.33 (1.85)	4.95* (1.69)	4.37 (2.03)	5.05* (2.04)

* The asterisk represents statistically significant figures at the 99% level.

In High Variability Phonetic Training, listeners are exposed to highly variable sound stimuli produced by multiple talkers in multiple phonetic contexts. Since learners can learn the relevant phonetic cues that apply to different listening situations, High Variability Phonetic Training has been proven to be very effective in improving not only learners' sound discrimination ability (cf. Lively et al., 1993; Logan et al., 1991) but also their pronunciation accuracy (Bradlow et al., 1995). In this training, the participants listened to English minimal pairs or triplets that differed only in a vowel sound or a consonant sound, such as '*feet-fit*' and '*tin-thin*'. These word pairs or triplets were presented both in isolation and embedded in sentences. 13 vowels and 17 consonants were included and each pair or triplet was made of phonemes that were thought to be confusable to Korean learners of English. In an attempt to assess the educational effect of our sentence stress feedback system, 20 participants were classified as the experiment group and they took part in additional English rhythm practice sessions using our online CALL system where the proposed sentence stress feedback system was incorporated. The 20 subjects in the control group were not given any conventional rhythm training because the purpose of this training experiment was not to prove that our method is superior to conventional teaching methods, but to observe whether the proposed system is effective in teaching English rhythm without teachers' instruction.

These training sessions were conducted 5 days a week, for 4 weeks. One High Variability Phonetic Training session took approximately 10 to 15 minutes and the average time required to complete one rhythm practice session with our system was 8.8 minutes. They also participated in a pretest and posttest during the first and last training session, respectively.

For the pretest and posttest, participants were asked to read the "Rainbow" passage, which is one of the most common standard reading passages used in speech evaluation tests, and their read speech was recorded. The recordings obtained from the pretest and posttest were randomized in order and presented to three native English speakers majoring in linguistics. They assessed the comprehensibility, accentedness, and rhythm of each participant's recordings on a 9 point scale, where 1 stood for 'impossible to understand or not at all native-like' and 9 stood for 'extremely easy to understand or native-like' (cf. Munro and Derwing, 1995 for an assessment of this type). The listeners were instructed to assess the speech rhythm of each participant based on sentence stress placement and the degree of stress-timing.

The results of the assessment in Table 5 suggest that only the participants in the experiment group, who used our sentence stress feedback system, showed significant improvements in rhythm and accentedness scores. Paired t-tests revealed that the accentedness and rhythm scores of each participant's post-test recording (accentedness: Mean: 4.95, SD: 1.69; rhythm: Mean: 5.05, SD: 2.04) were significantly different from those of their pre-test recording (accentedness: M: 4.33, SD: 1.85; rhythm: M: 4.37, SD: 2.03) in the experiment group (accentedness: $t(df = 59) = -3.43$, $p < .05$; rhythm: $t(df = 59) = -3.84$, $p < .001$), but not in the control group. Specifically, the participants in the experiment group received significantly higher accentedness and rhythm scores after the training, which suggests that their speech was rated as 'less foreign-accented' and 'having more native-like rhythm' after the training was received.

Additionally, the changes in the scores were converted into ratios to compare the degree of improvement between the two groups more precisely. The improvement ratio was calculated by subtracting a pre-test score from a post-test score, dividing the difference by the pre-test score, and then multiplying the ratio by 100, so that the value is converted into a percentage. As shown in Table 6, the improvement ratios of their accentedness and rhythm scores were 14.3% and 15.5%, respectively, whereas those of the participants in the control group were only 4.7% and 5.5%, respectively. This suggests that our system was successful in increasing learners' accentedness and rhythm scores by 10%.

This training experiment was conducted to evaluate the educational effect of our CALL system. The results demonstrate that learners who receive sentence stress training with our CALL system can improve their use of English rhythm to a greater extent than those who do not in comparison. However, their comprehensibility scores did not improve after the training. This result is not surprising because the correlation between comprehensibility and accentedness is not strong (Derwing and Munro, 1997; Munro and Derwing, 1995); comprehensibility is affected by a wider range

Table 6
Improvement ratio.

Criterion	Comprehensibility	Accentedness	Rhythm
Control group	4.6%	4.7%	5.5%
Experiment group	3.6%	14.3%	15.5%

of factors than accentedness, including those such as speech rate and word recognition (Jun and Li, 2010). Since our training focused on improving learners' English rhythm in a short period of time, it appears that our training was not sufficient to improve learners' overall spoken English performance. Nonetheless, the native listeners' ratings clearly show that our sentence stress feedback system helps non-native learners produce English rhythm in a more native-like way.

5. Discussion and conclusion

So far, we have described our unique sentence stress feedback system in which sentence stress prediction, detection, and feedback provision models are combined to offer feedback on sentence stress errors. The objective of the system is to provide non-native learners with corrective feedback so that they can improve their English rhythm and fluency.

The reference stress pattern of an input sentence is extracted from a stress prediction model trained with the Aix-MARSEC database. By using the Aix-MARSEC database, in which sentence stress is explicitly annotated, we were able to achieve an accuracy of 96.6% in sentence stress prediction. This accuracy is much higher than the accuracy of the pitch accent prediction model trained with the BURNC in our previous study (Kang et al., 2012).

To reflect the characteristic prosodic properties of Korean learners' English in the detection model, the KLEAC database was used to train the stress detection model; in the KLEAC database, the sentence stresses imposed by Korean learners are marked. As a result, the accuracy of the detection model was 84.1%. The proposed stress detection model detects the sentence stress of any input sentence with this accuracy whereas the detection model developed by Imoto et al. (2002) requires an existing reference utterance of a native speaker, which makes their system limited in the number of the utterances that can be input.

To automatically provide non-native learners with appropriate feedback on their sentence stress errors, a feedback provision model was constructed. This model was constructed based on the output deviations of the prediction and detection models and the optimum feedback decision boundaries that showed maximum correlation between learners' feedback scores and the KLEAC database speakers' proficiency levels. The decision boundaries θ_1 and θ_2 determined for the feedback provision model were 0.18 and 0.61, respectively.

In an experiment that evaluated the usefulness of the proposed sentence stress feedback system, we found that the students who trained with our system significantly improved their accentedness and rhythm, and improvement ratios were approximately 10% higher in the experiment than in the control group. Therefore, we believe that our sentence stress feedback system will be useful in helping non-native learners improve their English rhythm and fluency in a non-classroom environment.

The problem with our sentence stress feedback system is that it inevitably produces errors in sentence stress prediction, detection, and feedback because of the imperfection of sentence stress labels in the Aix-MARSEC and KLEAC databases as well as the POS tagger we used. We believe, however, that non-native learners will be more satisfied with the current version of this system if this system is combined with good learning contents in an appropriate way.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0008835). We are grateful to GLEC Institute for kindly helping us conduct the training experiment at their facility. We are also grateful to anonymous reviewers for their invaluable comments and to Sylvia Cho for proofreading the draft of this paper.

References

Abercrombie, D., 1964. Syllable quantity and enclitics in English. In: Abercrombie, D., Jones, D. (Eds.), In Honour of Daniel Jones. Longmans, London, pp. 216–222.

Abercrombie, D., 1967. Elements of General Phonetics. Edinburgh University Press, Edinburgh.

Anderson-Hsieh, J., Johnson, R., Koehler, K., 1992. The relationship between native speaker judgment of non-native pronunciation and deviance in segments, prosody, and syllable structure. Lang. Learn. 42, 529–555.

Bang, J., Kang, S., Lee, G., 2013. An automatic feedback system for English speaking integrating pronunciation and prosody assessments. In: Proceedings of SLaTE 2013. pp. 83–89.

Boersma, P., Weenink, D., 2009. Praat: doing phonetics by computer (Version 5.1.05) [Computer program], http://www.praat.org/.

Bolinger, D.L.M., 1958. A theory of pitch accent in English. Word 14 (2–3), 109–149.

Boula de Mareüil, P., Vieru-Dimulescu, B., 2006. The contribution of prosody to the perception of foreign accent. Phonetica 63 (4), 247–267.

Bradlow, A.R., Pisoni, D.B., Yamada, R.A., Tohkura, Y., 1995. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. J. Acoust. Soc. Am. 101 (4), 2299–2310.

Brill, E., 1992. A simple rule-based part of speech tagger. In: Proceedings of the third conference on Applied Natural Language Processing, Morristown. pp. 112–116.

Chandel, A., Parate, A., Madathingal, M., Pant, H., Rajput, N., Ikbal, S., et al., 2007. Sensei: spoken language assessment for call center agents. In: 2007 IEEE Workshop on Automatic Speech Recognition & Understanding. pp. 711–716.

Chen, L., Zechner, K., 2011. Applying rhythm features to automatically assess non-native speech. In: Proceedings of INTERSPEECH 2011. pp. 1861–1864.

Cheng, J., 2011. Automatic assessment of prosody in high-stakes English tests. In: Proceedings of INTERSPEECH 2011. pp. 1589–1592.

Dauer, R.M., 1983. Stress-timing and syllable-timing reanalyzed. J. Phon. 11, 51–62.

Derwing, T.M., Munro, M.J., 1997. Accent, intelligibility, and comprehensibility: evidence from four L1s. Stud. Second Lang. Acquis. 20 (1), 1–16.

Derwing, T.M., Rossiter, M.J., 2003. The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. Appl. Lang. Learn. 13 (1), 1–18.

Derwing, T.M., Munro, M.J., Wiebe, G., 1998. Evidence in favor of a broad framework for pronunciation instruction. Lang. Learn. 48 (3), 393–410.

Elliott, A.R., 1997. On the teaching and acquisition of pronunciation within a communicative approach. Hispania 80 (1), 95–108.

Gimson, A.C., 1980. An Introduction to the Pronunciation of English, third ed. Edward Arnold, London.

Hahn, L.D., 2004. Primary stress and intelligibility: research to motivate the teaching of suprasegmentals. TESOL Quart. 38 (2), 201–223.

Hirst, D., De Looze, C., Auran, C., Bouzon, C., 2009. Aix-MARSEC: a database for the analysis of the prosody of British English. [Database available from the Speech & Language Data Repository: http://sldr.org/sldr000033/en].

Hönig, F., Batliner, A., Weilhammer, K., Nöth, E., 2010. Automatic assessment of nonnative prosody for English as L2. In: Proceedings of Speech Prosody.

Imoto, K., Tsubota, Y., Raux, A., Kawahara, T., Dantsuji, M., 2002. Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system. In: Proceedings of ICSLP, Denver. pp. 749–752.

Ito, A., Konno, T., Ito, M., Makino, S., 2009. Evaluation of English intonation based on combination of multiple evaluation scores. In: Proceedings of INTERSPEECH 2009. pp. 596–599.

Jassem, W., 1952. Stress in modern English. Bull. de la Soc. Polonaise de Linguistique 12, 189–194.

Jassem, W., 1999. English stress, accent and intonation revisited. Speech and Lang. Technol. 3 33–50.

Jeon, J.H., Liu, Y., 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2009. Taipei, pp. 4565–4568.

Jones, D., 1972. An outline of English phonetics, ninth ed. Heffer, Cambridge.

Jun, H.G., Li, J., 2010. Factors in raters' perceptions of comprehensibility and accentedness. In: Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference. Iowa State University, pp. 53–66.

Kang, S., Lee, G.G., Lee, H.Y., Kim, B., 2012. An automatic pitch accent feedback system for English learners with adaptation of an English corpus spoken by Koreans. In: Proceedings of IEEE Workshop on Spoken Language Technology (SLT 2012), Miami. pp. 432–437.

Kingdon, D., 1958. The Groundwork of English Stress. Longmans, London.

Lafferty, J., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning 2001. pp. 282–289.

Lee, C., Jung, S., Kim, K., Lee, D., Lee, G.G., 2010. Recent approaches to dialog management for spoken dialog systems. J. Comput. Sci. Eng. 4 (1), 1–22.

Lee, H.Y., 2011. Evaluation of Korean Learners' English Accentuation. In: Proceedings of the 16th National Conference of the English Phonetic Society of Japan and the Second International Congress of Phoneticians of English. pp. 23–25.

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G.G., Sagong, S., et al., 2011. On the effectiveness of robot-assisted language learning. ReCALL J. 23 (1), 25–58.

Levy, M., 1997. CALL: Context and Conceptualisation. Oxford University Press, Oxford.

Liscombe, J., 2007. Prosody and speaker state: paralinguistics, pragmatics, and proficiency. PhD Dissertation, Columbia University.

Lively, S.E., Logan, J.S., Pisoni, D.B., 1993. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. J. Acoust. Soc. Am. 94 (3), 1242–1255.

Lloyd James, A., 1940. Speech Signals in Telephony. Pitman & Sons, London.

Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. J. Acoust. Soc. Am. 89 (2), 874–886.

Maier, A., Hönig, F., Zeißler, V., Batliner, A., Körner, E., Yamanaka, N., et al., 2009. A language-independent feature set for the automatic evaluation of prosody. In: Proceedings of INTERSPEECH 2009. pp. 600–603.

Mostow, J., Duong, M., 2009. Automated assessment of oral reading prosody. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009). pp. 189–196.

Munro, M.J., Derwing, T.M., 1995. Foreign accent, comprehensibility and intelligibility in the speech of second language learners. Lang. Learn. 45 (1), 73–97.

Pierrehumbert, J.B., 1980. The Phonology and Phonetics of English Intonation. Ph.D Dissertation, MIT.

Pike, K., 1945. The Intonation of American English. University of Michigan Press, Ann Arbor.

Qian, Y., Wu, Z., Ma, X., Soong, F., 2010. Automatic prosody prediction and detection with Conditional Random Field (CRF) models. In: Proceedings of ISCSLP 2010. pp. 135–138.

Rangarajan Sridhar, V., Bangalore, S., Narayanan, S.S., 2008. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Trans. Audio Speech Lang. Process. 16 (4), 797–811.

Roach, P., 1982. On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages. In: Crystal, D. (Ed.), Linguistic Controversies. Edward Arnold, London.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Pierrehumbert, J., Hirschberg, J., et al., 1992. TOBI: a standard scheme for labeling prosody. In: Proceedings of the International Conference on Spoken Language 92. Banff, pp. 563–566.

Sitaram, S., Mostow, J., Li, Y., Weinstein, A., Yen, D., Valeri, J., 2011. What visual feedback should a reading tutor give children on their oral reading prosody? In: Proceedings of SLaTE 2011. pp. 24–26.

Suzuki, M., Konno, T., Ito, A., Makino, S., 2008. Automatic evaluation system of English prosody based on word importance factor. J. Systemics Cybern. Inform. 6 (4), 83–90.

Tajima, K., Port, R., Dalby, J., 1997. Effects of temporal correction on intelligibility of foreign-accented English. J. Phon. 25 (1), 10–24.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K., Sönmez, K., 2000. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In: Proceedings of ICSLP. pp. 187–190.

Tepperman, J., Stanley, T., Hacioglu, K., Pellom, B., 2010. Testing suprasegmental English through parroting. In: Proceedings of Speech Prosody. p. 2010.

Wenk, B., Wiolland, F., 1982. Is French really syllable-timed? J. Phon. 10, 193–216.

Wennerstrom, A., 2000. The role of Intonation in Second Language Fluency. Perspectives on Fluency. University of Michigan Press, Ann Arbor, pp. 102–127.

Witt, S., Young, S.J., 1997. Language learning based on non-native speech recognition. In: Proceedings of ICASSP 1997. pp. 633–636.

Yamashita, Y., Kato, K., Nozawa, K., 2005. Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison. IEICE Trans. Inf. & Syst. E88-D (3), 496–501.

Zechner, K., Xi, X., Chen, L., 2011. Evaluating prosodic features for automated scoring of non-native read speech. In: Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 461–466.