LETTER

Reference-Independent Prosody Evaluation Based on Prosodic Unit Segmentation

Sixuan ZHAO[†a)], ***Student Member***, **Soo Ngee KOH**[†], ***and*** **Kang Kwong LUKE**[†], ***Nonmembers***

SUMMARY This paper proposes prosodic unit based segmentation for prosody evaluation by using pitch accent detection and forced alignment techniques. Support Vector Machine (SVM) is used to evaluate the prosody of non-native English speakers without reference utterances. Experimental results show the superiority of prosodic unit segmentation over word segmentation in terms of classification accuracy and dimension of the feature vectors used by SVM.

key words: *prosody evaluation, prosodic unit, pitch accent, segmentation*

1. Introduction

Most of the current prosody evaluation systems perform segmentation at the word or syllable level [1]–[3]. An input utterance is first segmented into words or syllables so that appropriate feature vectors can be extracted for prosody evaluation. Even though such segmentation methods work well for evaluating pronunciation, they may not be appropriate for applications that involve larger units such as phrases and sentences. Unlike units such as word and syllable which are based on a series of consonants and vowels, prosody is a supra-segmental feature and may not correlate with word boundaries. As a result, with segmentation based on lexical units, the evaluation results may fail to reflect the learner's mastery of prosody accurately. One logical solution is to consider the use of a suitable unit in the prosodic domain. According to prosodic theory, prosodic units reflect rhythm and phrasing skills rather than lexical and syntactic meaning as conveyed by lexical units. Thus, it is more reasonable to perform segmentation based on prosodic unit for prosody evaluation.

2. Prosodic Unit Segmentation

Foot is defined as a phonological unit consisting of an accented syllable followed by a series of unaccented syllables [4], [5]. It means that a foot always starts from the beginning of a stressed syllable to the beginning of the next stressed syllable. As an example, the sentence "*I felt that I might never stop the machine from running*" can be segmented at foot level as: "*/I /felt that I /might/ never /stop the ma/chine from/ running/*". In this sentence, one foot boundary does locate inside the word "machine". The initial unaccented phrase "*I*" (the first accented syllable is "felt") is called anacrusis from linguistics view point. However, it is also considered as one segmentation unit to model a complete prosody contour.

Foot is selected as the segmentation unit due to its appropriate length as well as its correlation with English rhythm and stress. First of all, it is obvious that the length of the segmentation unit should be limited; otherwise, problems may arise for short sentences. Foot possesses suitable length for segmentation and preserves enough prosodic information. Secondly, the definition of foot determines its influence on rhythm information of an utterance. As a stress-timed language [4], English possesses a tendency to keep the length of each foot within a limited distance from the norm which is determined by the tempo at each moment of an utterance. Hence, foot expresses significant rhythmic information and contributes to the evaluation results.

3. Automatic Segmentation

According to definition, foot correlates to stresses or pitch accents tightly. Previous work [6] on stress and pitch accent detection provides cues to automate the process of segmenting a sentence into feet. From [6], pitch accent detection at word-level using logistic regression can achieve the highest accuracy. Furthermore, as most of foot boundaries in an utterance are still correlated to the boundaries of stressed words (more than 90% in our experimental corpus), it is reasonable to perform quasi-foot segmentation based on accented (stressed) words to segment utterances using logistic regression:

$$f(z) = \frac{e^z}{e^z + 1} \tag{1}$$

$$z = w \bullet x \tag{2}$$

where x is the feature vector, w is the weight vector, and $f(z)$ is the output between 0 and 1, which can be used as the probability of accentuation.

The steps for quasi-foot segmentation are listed as follows. First, word level segmentation is obtained by forced alignment. Second, logistic regression is performed to detect pitch accent at the word level. As in [6], normalized pitch, energy and duration are used as the input feature vector. Pitch slope, calculated as the slope of the 1-st order pitch regression line across two neighboring frames, is also added as an extra feature. Pitch is extracted by the subharmonic-to-harmonic ratio method (SHR) [7], with an average estimation error of 5 Hz based on CSTR database as reported

Manuscript received February 26, 2013.
Manuscript revised May 6, 2013.

[†]The authors are with Nanyang Technological University, Singapore.
a) E-mail: zhao0120@e.ntu.edu.sg

DOI: 10.1587/transinf.E96.D.2143

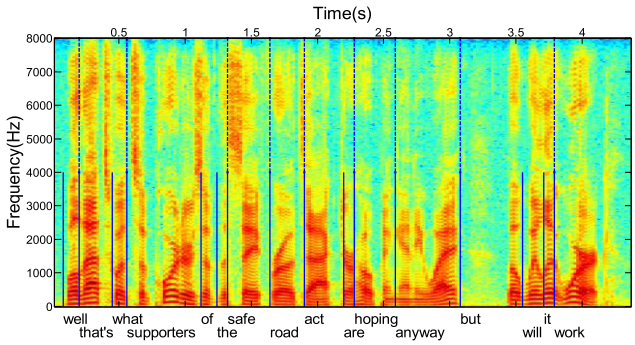

Fig. 1 Example of quasi-foot segmentation.

Table 1 Pitch accent and foot boundary detection results.

	Detection Method	Evaluation Reference	Detection Accuracy (F-measure)
Pitch Accent Detection	Logistic Regression	Compared with real pitch accent	0.86
Quasi-foot Segmentation	Forced Alignment +Logistic Regression	Compared with real foot boundaries	0.81

in [8]. Finally, the boundary of a detected accented word is taken as the quasi-foot boundary. In Fig. 1, the automatic quasi-foot segmentation is performed on the sentence "*Well/ that's what/ supporters of the/ Safe/ Road/ Act are/ hoping/ anyway*":

The upper solid lines are word boundaries obtained by forced alignment, whereas the upper dash lines are quasi-foot boundaries detected by the proposed method. All the foot boundaries in this example are the same as manually detected boundaries by experts. Besides, in Fig. 1, words such as "supporters", "Safe", "Road", "Act", "hoping" and "anyway" are all detected as accented words. This result is reasonable since human speakers also usually put stress on those kinds of words.

To analyze the performance of the quasi-foot segmentation method, detection experiments are performed on 60 utterances from the Boston University Radio News Corpus (BURNC). The detected boundaries are compared with manually labeled pitch accent and foot boundaries, and F-Measures are shown in Table 1:

As compared to real pitch accent, logistic regression based pitch accent detection yields an F-measure of 0.86. For the case of quasi-foot segmentation based on pitch accent detection and forced alignment, the resulting F-Measure is 0.81 which is lower than that for accent detection. Although not perfect, the boundaries detection is still reasonably good to be used for prosody evaluation.

4. Experiments and Discussions

A reference-independent evaluation model based on SVM is used to evaluate the prosody without using pre-recorded reference utterances. Three segmentation methods based on (1) foot, (2) word and (3) voiced/unvoiced (v/uv) as in [3] are used and the evaluation results are compared. The prosody evaluation is performed by SVM based regression /classification. To be consistent with previous works, e.g., [1]–[3], a 5-level score (with 1 as the worst and 5 as the best) is assigned to each sentence by human evaluators. After training the SVM based on mean subjective scores from human evaluators, the machine score can be estimated according to the regression/classification model and the input feature vector without using the reference utterances. In our experimental design, a 60-dimensional feature vector using information relevant to regression line, max., min., and positions (start, end, max., min.) of pitch and energy is used as in [3]. It should be noted that the main purpose of the letter is to discuss segmentation methods rather than feature extractions for prosody evaluation.

Table 2 Reference-independent prosody evaluation results.

Segmentation Methods	Human-machine correlation (regression)	Classification Accuracy (classification)
quasi-foot	0.61	51.7%
word	0.54	46.4%
v/uv	0.45	42.3%

A total of 200 utterances with 20 unique sentences from 10 different non-native English speakers (Chinese, Vietnamese and Indians) are collected as student utterances and a total of 60 utterances with the same transcriptions from BURNC are extracted and used as teachers' utterances. Three evaluators who are linguists and are native speakers of English assess the prosody of all the sentences, with an inter-evaluator correlation of 0.64. SVM models are implemented by LIBSVM [9], using a radial basis function (RBF) kernel.

The experimental results obtained are shown in Table 2. Classification accuracy refers to the percentage that the machine scores are same as the corresponding human scores using classification model. The human-machine correlation coefficients for the regression model and the inter-evaluator correlation coefficients are calculated based on Spearman's definition. T-test results show $p < 0.01$ for each pair of segmentations.

From the obtained results, it is clear that quasi-foot segmentation achieves the best results in terms of human-machine correlation and classification accuracy. The quasi-foot segmentation shows a human-machine correlation (0.61) which is close to the inter-evaluators correlation (0.64). In addition, this reference-independent evaluation scheme does not rely on pre-selected sentences and thus can evaluate any input utterances with the corresponding transcriptions given by the learner. Therefore, learners can practice on arbitrary sentences rather than being restricted by the reference corpus as in [1], [2].

To further examine the contribution of the proposed segmentation approach, the accented word ratio which is defined as the number of accented word divided by the total

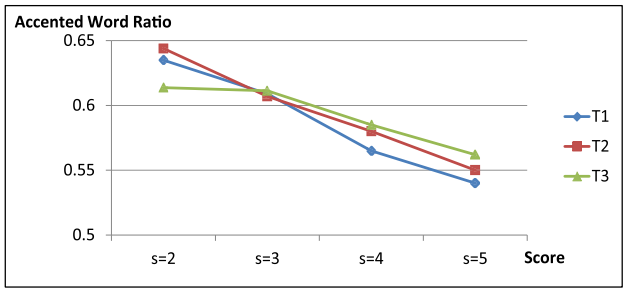

Fig. 2 Accented Word Ratios and Human Scores.

number of words in an utterance is calculated for each utterance with a score from 2 to 5. Here, score 1 is excluded because only very few utterances are scored as 1 by the evaluators. The mean accented word ratios for the test utterances with a score of 2 to 5 given by 3 evaluators (T1, T2 and T3) are plotted in Fig. 2:

It can be seen that the accented word ratio decreases with the increase in the subjective scores given by the evaluators. As non-native speakers with poorer speaking skills tend to accentuate most of the word in their attempt to make correct pronunciations, more accented word will be detected in their utterances. In contrast, speakers with proficient speaking skills can manipulate stresses well and accentuate the words appropriately according to the rhythm of the sentence, thus leading to fewer accented words. Word segmentation or voiced/unvoiced segmentation does not differentiate accented and unaccented word, thus the obtained feature vector will not be affected by rhythmic information. In contrast, quasi-foot segmentation is based on accented word detection and each segment correlates to one accented word. As a result, the rhythmic information is modeled and contributes to the prosody evaluation results.

5. Feature Selection Tests

In addition to the accented word ratio, a feature selection test can be performed to test the robustness of the proposed method. The minimum Redundancy Maximum Relevance (mRMR) method proposed by [10], which tries to minimize the redundancy between each pair of features and maximize the relevance between selected features and the class label, is adopted:

$$\max D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{3}$$

$$\min R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \tag{4}$$

where x_i is the i-th feature, c is the label, and I refers to the entropy between feature pairs or features and labels.

Figure 3 shows the relationship between the feature dimension and the human-machine correlation based on SVM regression. The quasi-foot based segmentation method always results in the highest human-machine correlation compared with the two other methods, regardless of the feature

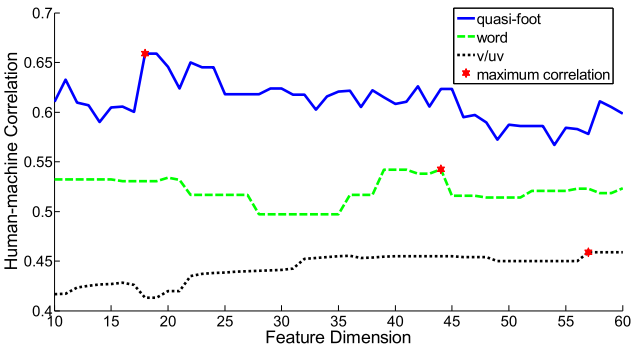

Fig. 3 Feature Dimension vs. Human-machine Correlation.

dimensions. Besides, the quasi-foot segmentation method achieves the best correlation coefficient 0.66 with a feature dimension of 18, which is smaller than that for word segmentation (44) and v/uv segmentation (57). It means that quasi-foot segmentation can obtain a reasonable accuracy with a comparatively smaller feature subset, leading to higher efficiency.

6. Conclusion

This letter proposes a new segmentation unit, namely foot, for computer-aided prosody evaluation and develops an automatic segmentation method for practical implementation. Experiments based on reference-independent models show that the proposed quasi-foot segmentation outperforms other segmentation methods in terms of classification accuracy and human-machine correlation. Feature selection experiments also show the superiority of the proposed segmentation method.

Acknowledgement

The authors would like to acknowledge the Ph.D. grant from the Institute for Media Innovation, Nanyang Technological University. The authors also would like to thank the anonymous reviewer for valuable comments.

References

[1] A. Ito, T. Konno, M. Ito et al., "Intonation evaluation of English utterances using synthesized speech for computer-assisted language learning," Int. J. Innovative Computing Information and Control, vol.6, no.3(B), pp.1501–1514, March 2010.
[2] M. Suzuki, T. Konno, A. Ito et al., "Automatic evaluation system of English prosody based on word importance factor," J. Systemics, Cybernetics and Informatics, vol.6, no.4, pp.83–90, 2008.
[3] A. Maier, F. Honig, V. Zeissler et al., "A language-independent feature set for the automatic evaluation of prosody," Proc. Interspeech, Brighton, UK, 2009, pp.616–619, 2009.
[4] A. Fox, Prosodic features and prosodic structure: The phonology of suprasegmentals, Oxford University Press, 2000.
[5] S. Zhao, K.K. Luke, S.N. Koh et al., "Computer aided evaluation of intonation for language learning based on prosodic unit segmentation," Proc. APSIPA ASC, 2010, pp.788–793, 2010.
[6] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the

word, syllable and vowel level," Proc. Human Language Technologies, 2009, pp.81–84, 2009.

[7] X.J. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," Proc. ICASSP, pp.333–336, Orlando, USA, 2002.

[8] I. Luengo, I. Saratxaga, E. Navas et al., "Evaluation of pitch detection algorithms under real conditions," Proc. ICASSP, Hawaii, USA, 2007, pp.1057–1060, 2007.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intelligent Systems and Technology (TIST), vol.2, no.3, pp.27:1–27:27, 2011.

[10] H.C. Peng, F.H. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.8, pp.1226–1238, Aug. 2005.