Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Assessment of L2 English Word Prosody Using Weighted Distances of F0 and Intensity Contours

Quy-Thao Truong[1], *Tsuneo Kato*[1], *Seiichi Yamamoto*[1]

[1] Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan

duq3104@mail4.doshisha.ac.jp, {tsukato, seyamamo}@mail.doshisha.ac.jp

Abstract

In the current paper, an automatic prosody assessment method for learners of English using a weighted comparison of fundamental frequency (F0) and intensity contours is proposed. Patterns of F0 and intensity of learners are compared to that of native using a proposed metric – a weighted distance – in which the error around the high values of prosodic features have more weight in the computation of the final distance. Gold-standard native references are built using the k-means clustering algorithm. Therefore, we also propose a data-driven criterion called weighted variance based on the weighted similarity within the whole set of native utterances to determine the optimal number of clusters k. In comparison with baseline contour comparison metrics which resulted in a subjective-objective score correlation of 0.278, our method combining the proposed metric and criterion led to a final subjective-objective score correlation of 0.304. In comparison, subjective scores correlated at 0.480.

Index Terms: prosody assessment, prosodic contour modeling, language learning

1. Introduction

It can be challenging for language learners to find an appropriate environment in which to practice the language, whether it is due to fear of talking in a traditional class or because human teachers can be costly or scarce. Over the past years, computer assisted pronunciation training (CAPT) systems have provided learners solutions to overcome these issues by offering technologies for correcting and giving feedback on the learners' pronunciation. Although most of these technologies have mainly focused on segmental aspects of speech, which handle phonetic pronunciation, attention has to be paid to suprasegmental aspects of speech responsible for intonation, rhythm and stress patterns, also known as prosody. Prosody plays a significant role in language learning since it can reflect a learner's fluency and intelligibility.

Research has been conducted to introduce prosody assessment for language learners. Escudero et al. [1] proposed to use tones and break indices (ToBI prosodic labels) [2] on non-native and native utterances. Mutual information is computed between the set of labels to determine the quantity of information shared between them and infer the quality of the non-native speaker's prosody. A more common approach to non-native prosody assessment is to compare acoustic features of prosody – fundamental frequency (F0), intensity, and duration – of a learner and a reference utterance. Arias et al. [3] developed an intonation assessment system in which non-native speakers were asked to repeat a sentence following an intonation pattern described with rise and fall labels. The F0 contour of the non-native speaker is then compared to that of a proficient speaker who had to follow the same intonation pattern using a frame-by-frame dynamic time warping alignment. Although the method showed promising results, it is underlined in [4] that using such prosodic labels can be challenging since it requires expert knowledge and suffers from low inter-rater agreement. Instead of using only one speaker as a reference, Schwanenflugel et al. [5] compares a child's F0 contour to the average of a set of adult speakers' contours to evaluate the child's prosody. Children had to utter sentences and F0 contours were constructed by averaging the F0 values over each word of the sentence. As a final reference contour, the authors took the average of several adult contours. However, they assumed that one ideal F0 pattern existed among the adult speakers, so they made sure that the final averaged reference was made up of adult contours that highly correlated with each other. Speakers whose F0 contours did not correlate well with those of other speakers' were discarded from the analysis and not taken into account in the final averaged contour. As such methods do not allow to cover the diversity of prosodic contour variations, Wang et al. [6] proposed to consider each individual native contour as a reference, but doing so can end up being computationally expensive and inefficient. Instead, they correlated F0 contours of non-native read speech against a single averaged contour of all native speakers, this time regardless of the various existing patterns among them. To further improve the representation of the diversity of prosodic contours, Cheng [7] proposed to create three clusters of similar F0 and intensity contours using the k-means algorithm to allow several prosodic reference representations for a same utterance. This time, the references to which the non-native contours were compared were non-native contours that were assigned a high prosodic score by human raters. Therefore, no native database was needed and all speakers involved in the study were recorded under the same conditions. Moreover, unlike the studies previously cited, Cheng did not correlate the contours together but used the Euclidian distance between them as a comparison metric. In addition to F0 and intensity features, phoneme duration information was also used to predict final prosodic scores, which allowed the final correlation between machine and human scores to be greater than the correlation between human scores.

Motivated by the encouraging results obtained in the previous studies, we attempt to bring improvements to existing frameworks of automatic prosody assessment by prosodic contour comparison. Fundamental research on prosody acknowledges the fact that peaks in prosodic features, especially in F0, induce the perception of prominence (i.e., stress or accent) [8, 9]. Such affirmation led us to believe that when dealing with prosody, acoustic measures do not bear the same importance at every point; thus, it is necessary to discriminate low and high values, the latter being more meaningful than the former. While previous research tended to treat acoustic measures equally at every point, we propose a novel metric, a weighted distance, that would allow high values of prosodic features to have more importance in the contour comparison task. Moreover, it is cru-

10.21437/Interspeech.2018-1386

cial to have enough reference contours covering the diversity of possible prosodic patterns, but none of the previous studies provided a way to determine the appropriate number of references needed. As a consequence, we also propose a criterion called weighted variance based on the similarity between the set of individual contours to determine how many clusters are necessary when using the k-means algorithm to cluster the contours. Because obtaining accurate phoneme and syllable segmentations of non-native speech is out of the scope of the current paper, we decided not to include phoneme duration information in the analysis and only consider F0 and intensity features. Our method is thus independent of any speech recognition system so that the phonetic quality of the learner's utterance does not impede their prosody assessment. In the present paper, we focus on Japanese learners of English, but the method proposed can be extended to learners with various native languages.

2. Prosodic contour comparison

2.1. F0 and intensity feature processing

Prosodic contours, consisting of F0 and intensity contours, of each native and non-native responses have been extracted following the method proposed by Cheng in [7], in which F0 and intensity values are sampled at $N_t = 25$ equally spaced points throughout the utterance. Feature extraction is conducted using Praat intensity and pitch tracker [10]. To avoid the natural variations between speakers, all features are z-normalised at the speaker level with respect to corresponding mean and standard deviation values. However, standard deviation values can vary a lot depending on the number of utterances spoken by the same speaker. As a consequence, z-normalisation is likely to introduce outliers around the extreme values of the contours and impact the comparison between them. To address this issue, we suggest adding a step to the processing procedure which consists in normalising the prosodic contours with the following sigmoid function:

$$u(t) = \frac{1}{1 + \exp(-\alpha z(t))} \tag{1}$$

where $z(t)$ refers to the z-normalised value of the prosodic feature at time t and α represents the slope of the sigmoid function. Introducing this step allows the smoothing of quick variations around high values while conserving the general shape of the contour and keeping the features in a range of 0 to 1 for all speakers. In the rest of the paper, a prosodic contour is denoted as U, such that $U = (u(1), \ldots, u(N_t))$ is the concatenation of the N_t normalised prosodic values.

As stated in previous studies, peaks in prosodic values surrounded by valleys matter when identifying prominent syllables in a word [8, 9, 11]; therefore, we decided not to interpolate unvoiced regions where F0 values are undetermined but to keep them at a zero value so that the peak and valley characteristics found in raw F0 contours are preserved in the normalised ones.

2.2. Comparison metric

Common metrics for contour comparison between non-native and reference utterances include Euclidian distance [7], dynamic time warping [3] and Pearson correlation coefficient [5, 6]. To have high values of prosodic contours matter more in the evaluation of prosody, we propose a metric, referred to as "weighted distance" in the rest of the paper, which consists in putting more weight on the squared error between the reference and the non-native contour around high values of the reference F0 or intensity contour. The weighted distance $wDist$ between a sigmoid-normalised reference contour U_r and non-native one U_{nn} of a given word is defined as

$$wDist(U_r, U_{nn}) = \sum_{t=1}^{N_t} wErr(u_r(t), u_{nn}(t)) \tag{2}$$

where $wErr(u_r(t), u_{nn}(t))$ represents the weighted error between $u_r(t)$ and $u_{nn}(t)$, the feature value at time t of the reference and non-native contour, respectively. The weighted error is calculated as follows

$$wErr(u_r(t), u_{nn}(t)) = w_t(u_r(t)) \cdot (u_r(t) - u_{nn}(t))^2 \tag{3}$$

The weight $w_t(u_r(t))$ is linearly dependent on $u_r(t)$, where the highest value of U_r over the N_t sample points is assigned the highest weight and the lowest value is assigned the lowest weight

$$w_t(u_r(t)) = w_{min} + \frac{(u_r(t) - U_{rmin}))(w_{max} - w_{min})}{(U_{rmax} - U_{rmin})} \tag{4}$$

The minimal and maximal values, w_{min} and w_{max}, between which the weights vary will be tuned through experiment.

For each utterance to score, the proposed weighted distance between its F0 (resp. intensity) and a reference F0 (resp. intensity) contour is calculated. The final automatic score for the utterance to assess is the average distance between the one given for F0 contour comparison and that given for intensity contour comparison.

2.3. Gold-standard native references

Because different pronunciations of the same word with a good prosody can result in very different prosodic patterns [12], it is necessary to have several gold standard references at our disposal when it comes to comparing prosodic contours in the task of prosody assessment. We chose to use the k-means algorithm to generate several reference contours by clustering similar prosodic contours together, as suggested by Cheng [7]. Since optimal clusters in k-means should be able to describe the overall variability in the data while avoiding redundancy between the clusters, we propose a systematic criterion to set the *k* value based on how similar the contours are to each other – that is, the more similar they are, the smaller *k* will be. In consistency with our previously defined weighted distance (2), the proposed criterion is derived from the variance between contours, where the average Euclidian distance between the contours is replaced by our weighted distance. This criterion is referred to as the weighted variance ($wVar$) between the contours and is defined as:

$$wVar = \frac{1}{N} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} wDist({U_r}^{(i)}, {U_r}^{(j)}) \tag{5}$$

where

- $N = \binom{n}{2}$ is the number of pairs of contours in the total set of n contours
- $wDist({U_r}^{(i)}, {U_r}^{(j)})$ is the weighted distance, defined by (2), between two reference contours ${U_r}^{(i)}$ and ${U_r}^{(j)}$ of the same word

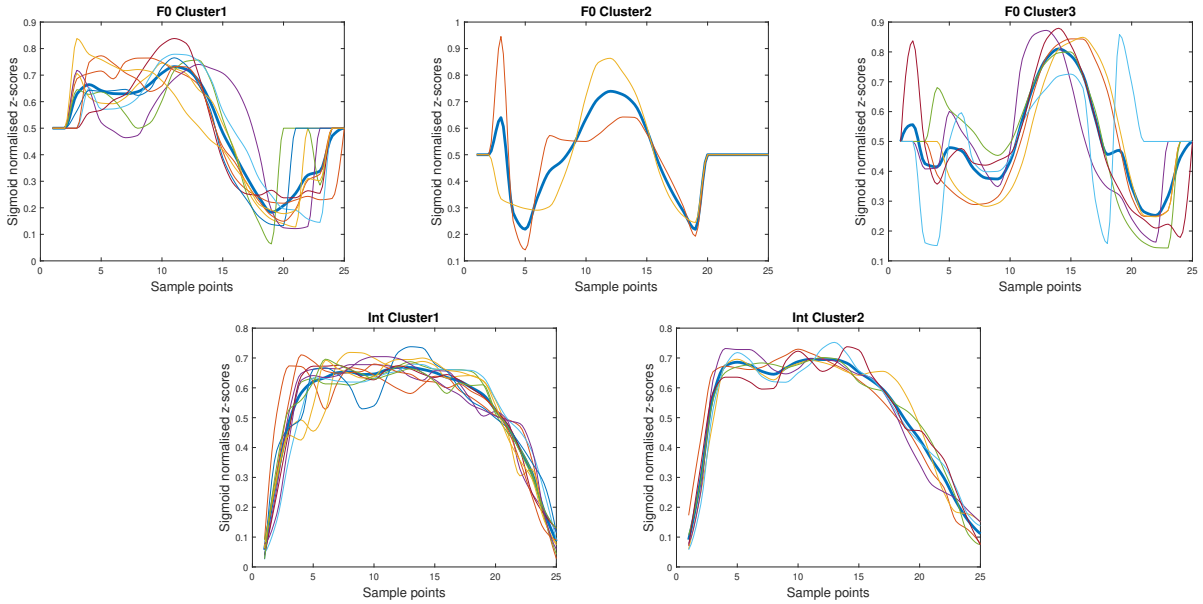

Figure 1: *Example of k-means results for the word "gorilla" on native utterances where k=3 for F0 contours (top) and k=2 for intensity contours (bottom) using our weighted variance criterion. Z-scores of prosodic contours have been converted by the sigmoid normalisation function defined in (1). Bold lines represent the mean contour of each cluster*

For each word, the weighted variance of the set of native prosodic contours is calculated and compared to the minimal and maximal weighted distances between two different contours of the set. We decided to make the number of clusters *k* linearly dependent on these minimal and maximal values as follows:

$$k = \lfloor k_{min} + \frac{(wVar - wVar_{min})(k_{max} - k_{min})}{(wVar_{max} - wVar_{min})} \rfloor \quad (6)$$

with

$$\begin{cases} k_{min} = 1 \\ k_{max} = \lfloor \frac{n}{2} \rfloor \\ wVar_{min} = \min\limits_{i,j \in [1,n], i \neq j} wDist(U_r{}^{(i)}, U_r{}^{(j)}) \\ wVar_{max} = \max\limits_{i,j \in [1,n], i \neq j} wDist(U_r{}^{(i)}, U_r{}^{(j)}) \end{cases} \quad (7)$$

If the weighted variance between the contours is small enough, we allow the k-means algorithm to create only one cluster ($k_{min} = 1$) so that the resulting reference contour corresponds to the average of all the contours. Moreover, in order to avoid at best the probability of creating empty clusters, the upper limit for the number of clusters k_{max}, is set set to half of the total number of reference contours.

Since the variance is computed independently for F0 and intensity contours of the same word, a different number of clusters can be generated for the two features. An example of clustering using our criterion is shown in Figure 1, where k-means is applied on one of the words of our native set of utterances decribed later in Section 3. In that example, a different number of clusters is generated for F0 and intensity contours.

In cases where several clusters are produced, the contours of the utterance to assess will be compared to the mean of each cluster. The value of the metric obtained for the closest mean will be kept to compute the final assessment score.

3. Experiments

3.1. Data

Experiments were conducted on isolated English words spoken by Japanese learners of English from the English Read by Japanese Corpus (ERJ) [13]. Among all the utterances of the corpus, 910 were prosodically assessed by two native American English teachers who had to evaluate the prosody of the speakers on a scale of 1 to 5, corresponding to categories ranging from "very poor" to "excellent". The raters were specifically asked to focus on whether or not the speaker positioned the stress on the appropriate syllable. This subset corresponds to 36 words with different numbers of syllables and various accent patterns spoken by 160 (79 female and 81 male) Japanese university students. Speakers with a wide range of proficiency in English were chosen for the development of the dataset. Examples of words found in the dataset include, for instance, dessert, totalitarian, percent, accessory and kangaroo. Overall, the subjective score correlation, defined as the correlation between the native raters, was 0.480. It should be pointed out that this low overall subjective score can be accounted for the fact that the scores of only two human raters were used in the study and that there is a relatively small number of utterances per word. Small variations in human ratings can then result in significant degradation of the subjective score correlation. This subjective score is nonetheless considered as an upper limit for final subjective-objective score correlations.

Native reference utterances for each word present in the dataset were taken from online English dictionaries [14, 15, 16, 17, 18, 19]. Depending on the availability of native pronunciations online, from 4 to 19 native utterances could be recorded with an average of 14 utterances per word. Speakers with various English accents were available (Australian, Irish, Jamaican, Scottish, UK, UK Received Pronunciation, UK Yorkshire, US, and US Southern).

3.2. Experimental conditions

To evaluate the performance of the proposed metric and gold-standard native reference building method, correlations between subjective-objective scores (that is, between human assigned scores and automatic scores) were computed and compared to those obtained with the baseline methods. Results are summarized in Table 1. The baseline metrics refer to the Euclidian distance (*euclidDist*) and the feature correlation (*featCorr*) between the contours. The baseline methods for building gold-standard native references refer to when all native contours are averaged into a single native reference (*mean*) as well as when the k-means clustering algorithm is conducted with a fixed number of clusters (*kMeans*, *k=4*). The first four lines of Table 1 denoted as (A), (B), (C) and (D) represent those baseline conditions.

The performance of our proposed weighted variance criterion (*wVar*) for choosing *k* in k-means was evaluated when used together with the baseline contour comparison metrics as well as with our weighted distance (conditions (E), (F) and (I) of Table 1). The performance of the baseline methods for building gold-standard references were also evaluated when used with our weighted distance (conditions (G) and (H)).

Results were obtained with tuned parameters such as the slope α of the sigmoid function from equation (1) ($\alpha = 1$) and the limit weights w_{min} and w_{max} from equation (4) ($w_{min} = 0.5$, $w_{max} = 1$).

Experiments were iterated 200 times and the final subjective-objective score correlations were averaged over the iterations using Fisher's method.

Table 1: *Subjective-objective score correlations*

Conditions	*Gold standard references*		*Contour comparison metrics*			Correlation
	mean	*kMeans*	*euclidDist*	*featCorr*	*wDist*	
(A)	•		•			0.250
(B)		*k=4*	•			0.265
(C)	•			•		0.277
(D)		*k=4*		•		**0.278**
(E)		*wVar*	•			0.279
(F)		*wVar*		•		**0.300**
(G)	•				•	0.286
(H)		*k=4*			•	0.294
(I)		*wVar*			•	**0.304**

The best performance on the baseline methods was obtained when the feature correlation was used as metric for contour comparison when gold-standard native references are built with *k=4* clusters (condition (D)). The feature correlation allowed to predict final prosodic scores even better when used with our weighted variance criterion (condition (F)). Regardless of the method used for building gold standard references, our weighted distance achieved better results than the baseline metrics (conditions (G), (H), (I)) but the best result was obtained when it was used together with our *wVar* criterion (condition (I)). The highest correlation between subjective-objective scores thus obtained was 0.304. As a comparison, the overall correlation subjective score of our dataset was 0.480.

4. Conclusions

We proposed a metric to determine the similarity between non-native utterances prosodic contours and gold-standard native reference contours to evaluate the quality of their speaker's prosody. Our proposed weighted distance allowed to outperform traditional contour comparison metrics such as the Euclidian distance or the feature correlation between the contours. Gold-standard references constructed with our weighted variance criterion to choose the number of clusters in the k-means algorithm also allowed better results than baseline methods to build gold-standard native references. The weighted variance enabled to create an appropriate number of *k* clusters so that the diversity of prosodic patterns was represented by the resulting *k* native references. The proposed metric and criterion took into account the theory that peaks of prosodic features account for the perception of prominence by assigning more weight to the error around peaks of prosodic values [8, 9]. Given the results obtained, we are encouraged to find further ways to include the prosodic theory previously cited into prosody assessment tasks. In that sense, the incorporation of additional features derived from F0 and intensity to help the discrimination of high and low values of prosodic features is a direction for future work. Moreover, we focused here on isolated English words but believe that the methods proposed can be extended to the study of sentences' prosody if used in combination with adequate prosodic feature processing such as the one proposed by Wang et al. [6].

5. References

[1] D. Escudero, C. Gonzalez-Ferreras, L. Aguilar, and E. Estebas, "Automatic assessment of non-native prosody by measuring distances on prosodic label sequences," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings*, 2017, pp. 1442–1446.

[2] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *ICSLP 1992 – 2nd International Conference on Spoken Language Processing, October 12-16, Banff, Alberta, Canada, Proceedings*, 1992, pp. 867–870.

[3] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Communication*, vol. 52, no. 3, pp. 254–267, 2010.

[4] J. Tepperman and S. Narayanan, "Better nonnative intonation scores through prosodic theory," in *INTERSPEECH 2008 – 9th Annual Conference of the International Speech Communication Association, September 22-26, Brisbane, Australia, Proceedings*, 2008, pp. 1813–1816.

[5] P. J. Schwanenflugel, A. Hamilton, J. Wisenbaker, M. Kuhn, and S. Stahl, "Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers," *Journal of Educational Psychology*, vol. 96, no. 1, pp. 119–129, 2004.

[6] X. Wang, K. Evanini, and S.-Y. Yoon, "Word-level F0 modeling in the automated assessment of non-native read speech," in *SLaTE 2015 – 6th Workshop On Speech and Language Technology in Education, September 4-5, Leipzig, Germany, Proceedings*, 2015, pp. 23–27.

[7] J. Cheng, "Automatic assessment of prosody in high-stakes english tests," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 27-31, Florence, Italy, Proceedings*, 2011, pp. 1589–1592.

[8] C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 3009–3022, 1997.

[9] A. Rietveld and C. Gussenhoven, "On the relation between pitch excursion size and prominence," *Journal of Phonetics*, vol. 13, no. 3, pp. 299–308, 1985.

[10] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[11] H. Mixdorff and O. Niebuhr, "The influence of F0 contour continuity on prominence perception," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25-29, Lyon, France, Proceedings*, 2013, pp. 230–234.

[12] M. Chu, Y. Zhao, and E. Chang, "Modeling stylized invariance and local variability of prosody in text-to-seeh synthesis," *Speech communication*, vol. 48, no. 6, pp. 716–726, 2006.

[13] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "English speech database read by japanese learners for call system development," in *LREC 2002 – 3rd International Conference on Language Resources and Evaluation, May 29-31, Las Palmas, Canary Islands - Spain, Proceedings*, 2002, pp. 896–903.

[14] *WordReference*, Accessed: 2018-02-02. [Online]. Available: https://www.wordreference.com/

[15] *Cambridge Dictionary*, Accessed: 2018-02-02. [Online]. Available: https://dictionary.cambridge.org/

[16] *Collins English Dictionary*, Accessed: 2018-02-02. [Online]. Available: https://www.collinsdictionary.com/dictionary/english

[17] *Macmillan Dictionary*, Accessed: 2018-02-02. [Online]. Available: https://www.macmillandictionary.com/

[18] *Oxford Dictionaries*, Accessed: 2018-02-02. [Online]. Available: https://en.oxforddictionaries.com/

[19] *Wiktionary, the free dictionary*, Accessed: 2018-02-02. [Online]. Available: https://en.wiktionary.org/