Evaluating Prosodic Features for Automated Scoring of Non-Native Read Speech

Klaus Zechner, Xiaoming Xi, Lei Chen

Educational Testing Service

Princeton, NJ, USA

{kzechner,xxi,lchen}@ets.org

Abstract—We evaluate two types of prosodic features utilizing automatically generated stress and tone labels for non-native read speech in terms of their applicability for automated speech scoring. Both types of features have not been used in the context of automated scoring of non-native read speech to date.

In our first experiment, we compute features based on a positional match between automatically identified stress and tone labels for 741 non-native read text passages with a human gold standard on the same texts read by a native speaker. Pearson correlations of up to r=0.54 between these features and human proficiency scores are observed.

In our second experiment, we use stress and tone labels of the same non-native read speech corpus to compute derived features of rhythm and relative frequencies, which then again are correlated with human proficiency scores. Pearson correlations of up to r=-0.38 are observed.

I. Introduction

A. Background and Related Work

When assessing the proficiency of non-native speakers in reading passages of connected text, the following four major dimensions are traditionally considered to be most relevant [1][2][3]: (1) reading accuracy, i.e., considering reading errors on the word level such as insertions, deletions or substitutions, compared to the reference text; (2) fluency, i.e., is the passage well paced in terms of speaking rate and distribution of pauses, and free of disfluencies such as fillers or repetitions; (3) pronunciation, i.e., are the words pronounced correctly on a segmental level (usually evaluated on individual phones); and (4) prosody, i.e., does the distribution of stressed and unstressed syllables, as well as the pitch contours of phrases and clauses match those of a native speaker.

While accuracy, fluency, and pronunciation measures have been explored in much detail in the past [3][4], this is much less the case for prosodic features. The reasons may include (1) that it is much harder and time-consuming to establish “references” or gold standards than for, e.g., pronunciation or fluency features; and (2) that prosodic features are much harder to compute reliably, in particular for non-native speech.

One notable exception, though, is the work by [5] who used machine-generated stress and tone labels for computing rhythm features in a non-native speech corpus. However, this work only looked at spontaneous speech. The second experiment of this paper will follow the spirit of this work quite closely, but now applied to a corpus of non-native read speech.

In other related research, rhythm metrics that do not require knowing stress or tone information have been used, [6][7]. In one study involving English speakers with Korean as their first language, it was found that the metrics regarding vocalic intervals are more effective than consonantal interval measures [6]. Another study reports on research on using such rhythm features for judging the nativeness of Japanese learners’ parroting of English [7].

While these rhythm features have the advantage of not relying on human annotations of stress and tone, we decided to base the prosodic features in this paper on stress and tone labels since stress and intonation are explicitly mentioned as scoring criteria in the guidelines for human raters of the read speech test responses and therefore, the precision of these labels’ location needs to be measured in automated speech scoring (first experiment in this paper).

B. Overview

In this paper, we look at the extent to which automatically predicted stress and tone labels can be used to compute prosodic features in an assessment for read speech, the ERT (English Reading Test). The current system for ERT is capable of computing features related to reading accuracy, fluency, and pronunciation, but still needs to be expanded into the prosody domain.

For the purpose of automatically generating stress and boundary tone information (boundary tones or just “tones” refer to the ending syllables of intonational phrases), we first develop decision tree classifiers for stress and tone labels trained on approximately 12,000 human annotated syllables of a non-native speech corpus. Then, we conduct two experiments of features utilizing these automatically generated stress and tone labels on a corpus of 741 text passages read by non-native speakers of English.

In the first experiment, the goal is to generate features that represent how well non-native speakers’ stress and tone patterns (as predicted by a classifier) match a human gold standard based on native speech. For this purpose, we automatically align time stamps of machine-predicted stress and tone labels with a reference gold standard, created by human expert annotators, and then correlate the obtained precision, recall, and F1-scores with human proficiency scores of the same text passages.

In the second experiment, the goal is to establish which measures of rhythm and frequency, derived from stress and tone patterns in non-native read speech, are most indicative of speech proficiency. For this purpose, we compute 12 features derived from machine-predicted stress and tone labels, indicating rhythm, stress and tone distributions, and their relative frequencies. Again, these features are correlated with human proficiency scores for the same text passages.

The remainder of this paper is organized as follows: Section II introduces the data we use for our study; Section III describes the automatic speech recognition (ASR) system and our system for automatic prediction of stress and tone labels for non-native speech; Section IV describes the two experiments to evaluate the usefulness of these prosodic features for automated speech scoring; Section V discusses the results and Section VI concludes the paper.

II. DATA

A. *Stress and Tone Labeled Data for Classifier Training*

To build the classifiers for predicting stress and tone labels, we used a corpus of 87 human-transcribed non-native spoken responses of about a minute in length each. The responses were annotated for stress and tone labels for each syllable by a native speaker of English. Since the development of these classifiers was done in the context of a speech scoring system targeted at spontaneous speech, this data was also drawn from a spontaneous speech corpus. Despite this mismatch in speaking ,mode between the training data for the classifiers and our test data (spontaneous vs. read speech), our results show that significant correlations for prosodic features can be obtained nonetheless. In future work, we will compare these results with those obtained from classifiers based on annotated read speech, i.e., in a matched condition scenario.

Before the human annotation process, forced alignment was used to obtain word and phoneme time stamps. The annotators used the Praat toolkit [8] for the annotation, which allowed them to listen to the audio sample, to look at its time and spectral representation and to enter label information. Following the ToBI schema [9], four tone labels as well as "no tone" were used. For stress, a binary scheme was used: stressed vs. unstressed syllable. Stressed syllables were defined as bearing the most emphasis or "weight" within a clause or sentence; typically, they coincide with lexical stress, but function words such as determiners or prepositions, as well as some content words may not bear stress.

Our data set had 28.1% stressed syllables and 12.9% syllables bearing a tone.

B. *Read Speech Data*

For the two evaluation experiments in Section IV, we used a corpus of 741 text passages, read by a total of 564 distinct non-native speakers of English; 387 speakers read one passage total, whereas 177 speakers read two passages each. The passages were drawn from three unique reading items in ERT administrations (item 1, item 2, item 3).

Further, the same three unique text passages were read by a native speaker of English and then annotated for stress and tone for each syllable by two native speakers of English who also were test development experts. They created a gold standard after initial annotations by adjudicating syllable labels with divergent annotations. We used the same binary scheme for stress annotation as above, but a simplified scheme for tone annotation that distinguishes only between rising and falling tone and "no tone".

Finally, we used forced alignment to map the annotated stress and tone labels of the human gold standard to the syllable time stamps of the corresponding text passages read by the native speaker.

Table 1 shows the distribution of stress and tone labels in this gold standard corpus based on native speech. We note that while the relative frequency of stressed syllables in the gold standard corpus is almost 10% higher than in the corpus used for decision tree training, the overall percentage of syllables with tone is by more than 5% lower. We conjecture that as a consequence of these distributional differences, the stress labels may be under-estimated, while the tone labels may be over-estimated by the classifiers on the read speech data.

Table 1. Distribution of stress and tone labels in the gold standard read speech corpus. All percentages are computed based on the number of syllables.

		Item 1	Item 2	Item 3	Total
	#words	96	58	60	214
	#syllables	160	92	91	343
Stress labels	Stressed	34.%	41.%	40.%	37.%
	Unstressed	65.%	58.%	59.%	62.%
Tone labels	Rising tone	0.6%	2.2%	3.3%	1.7%
	Falling tone	5.6%	5.4%	5.5%	5.5%
	No tone	93.%	92.%	91.%	92.%

III. AUTOMATIC STRESS AND TONE PREDICTION

A. *ASR System*

Our ASR system is a gender-independent continuous-density Hidden Markov Model (HMM) speech recognizer, initially trained on about 30 hours of non-native spontaneous speech, using additional spoken corpora for the language model [10]. To adapt the acoustic model for read speech, we used maximum a-posteriori (MAP) adaptation with 870 non-native read-aloud responses from ERT (about 12 hours of speech, disjoint from the data set containing 741 read passages used for the evaluations in Section IV).[1] We further built an

[1] This adaptation set contains passages from 6 ERT items, containing the 3 items used for the experiments in Section 4.

interpolated trigram language model (LM) where 90% of the weight is assigned to the LM built on the same set of 870 read-aloud passages and 10% to the initial LM of the recognizer (background LM). The word error rate (WER) of the adapted ASR system was 15.5% measured on an independent held-out evaluation set of 100 read-aloud passages (no speaker overlap with the adaptation set).

For the experiments in Section IV, our system ran in 2-pass mode, where it decoded the responses first using the default AM, trained on non-native speech, and then performed forced alignment with the hypotheses from Pass 1 using an AM trained on mostly native speech (see [11] for more details). The main reason for this set-up is to allow for maximum word accuracy in decoding of the non-native speech, while at the same time allowing the use of native speech acoustic characteristics for computing both pronunciation and prosodic features based on the forced alignment pass.

B. Features for Stress and Tone Prediction

Our system automatically extracted about 30 features based on power, pitch, duration, word-identity, syllable position within a word, dictionary stress, distance from last syllable with stress or tone, and pauses for every syllable (see [5] for a more detailed description). Furthermore, a context of five syllables prior and after the current syllable was also encoded for most of the features.

A total of 270 features were used as input features for classifier training (comprising the current syllable and the left and right contexts.)

C. Stress and Tone Classifiers

We trained decision tree (C4.5, [12]) classifiers to predict stress and tone information for every syllable, using a five-fold cross-validation set-up due to the rather small data size.

Tables 2 and 3 provide the results of these cross-validation experiments.

Table 2. Results of C4.5 classifier for stress prediction on a syllable level (classification accuracy and F1-scores).

Number of syllables	Accuracy	Stressed (F1)	Unstressed (F1)
12203	84.4%	69.6	89.5

Table 3. Results of C4.5 classifier for tone prediction on a syllable level (classification accuracy and F1-scores).

Number of syllables	Accuracy	Tone (F1)	No tone (F1)
12203	93.2%	64.5	96.2

We can observe that for stress, the classification accuracy of 84.4% is by12.5% absolute higher than the baseline of 71.9% (when marking all syllables as “unstressed”). For tone, the tone classification accuracy of 93.2% is by 6.1% higher than the baseline of 87.1%. In terms of relative error reduction, the error rate (100.0 - accuracy) on stress classification is reduced by 44.5% relative compared to the baseline, and for tone, by 47.3% relative compared to the baseline.

IV. Prosodic Feature Evaluation Experiments

A. Features Based on Comparing Predicted Labels with Human Gold Standard

In the first experiment, our system first generated stress and tone labels for 741 read-aloud passages based on the C4.5 decision trees described in the previous section. Since the distribution of tone labels in both the annotated data used for decision tree training (four tone labels) as well as in the human gold standard for native read speech (two tone labels) was highly skewed, we mapped all tone labels to a single label and thus achieve a binary classification, analogous to the stress classification.

The labels, along with their time stamps, were then converted into NIST’s[2] RTTM[3] format [13], with every vowel (syllable nucleus) appearing as a LEXEME line for basic alignment (time interval), and every stress or tone point (in separate files) as an IP [4] line (points in time). The corresponding human gold standard for the respective passage was then transformed into a time-warped reference file, where the total reading time is “stretched” (or “warped”) so that it matches the reading time of the non-native speaker (the average warping factor is about 1.34 which means that non-native speakers take on average about 34% longer to read the same passage than the gold standard native speaker).

NIST’s RTTM alignment and evaluation scripts [5] then generated statistics on differences of all 1482 pairs of references (i.e., gold standard labels) and hypotheses (i.e., automatically predicted labels) (741 files each for stress and tone), and we extracted information about precision, recall, and F1-score from the output files. We used a tolerance window of T=200msec to count stress or tone labels as correct match. Since the warping factor is likely to vary across the entire response, an even higher T might be warranted, but we chose a conservative approach to avoid over-estimation of system performance.

Tables 4 and 5 report correlations between human scores and precision, recall and F1-scores for the three passages (items 1, 2 and 3) as well as the entire set (ALL). The human scores are summations of two sub-scores, range from 0 to 6, and reflect accuracy and appropriateness in pronunciation, intonation, stress and pacing. We want to emphasize that these

[2] National Institute for Standards and Technology

[3] Rich TranscriptionTtime Markers. The RTTM format was used previously for meta-data research, including speech units, discourse markers, and disfluencies.

[4] IP stands for “interruption point” in the context of disfluency detection.

[5] We mainly used version 17 of md-eval.pl from SCTK.

tables do *not* report precision, recall and F1-scores directly (based on the match between predicted labels and gold standard labels), but rather their *correlations* with human proficiency scores. The results will be discussed in Section 5.

B. Features Derived from Predicted Labels

Our second experiment is looking at correlations with human scores of a set of 12 features that are derived from the automated stress and tone predictions. 10 of them can be seen as capturing aspects of global speech rhythm in terms of the pacing and evenness of stress and tones (features 1-10 in Table 6); the 2 other features indicate the relative frequency of stress and tone labels in a speaker's response (features 11 and 12 in Table 6).

Table 6 provides correlations between these 12 features derived from stress and tone predictions and human rater scores on the same data set as above (741 passages) and also provides definitions of the features.

Table 4. Correlations between precision, recall, and F1-scores of stress predictions and human rater scores for different read-aloud passages (items).

Item(s)	Correlations with Precision	Correlations with Recall	Correlation with F1scores
ALL (N=741)	**0.402**	**0.284**	**0.419**
1 (N=385)	0.333	0.111	0.317
2 (N=179)	0.543	0.338	0.537
3 (N=177)	0.451	0.311	0.429

Table 5. Correlations between precision, recall, and F1-scores of tone predictions with human rater scores for different read-aloud passages (items).

Item(s)	Correlation with Precision	Correlation with Recall	Correlation w.F1scores
ALL (N=741)	**0.276**	**0.306**	**0.305**
1 (N=385)	0.140	0.161	0.154
2 (N=179)	0.198	0.270	0.251
3 (N=177)	0.316	0.308	0.332

V. Discussion

In our first experiment, where we computed features based on the match in time location between predicted labels and human gold standard labels, we found an average correlation with human proficiency scores of r=0.42 for stress labels and r=0.31 for tone labels. In both instances, performance across different test items was quite variable; e.g., for item 2, the correlation of F1-scores and human proficiency scores reached r=0.54. This is higher than our best performing non-prosodic feature on this data set (r=0.503 for speaking rate).

Furthermore, we observe that for stress-label based features, while correlations between precision values and human scores are roughly at par with those of F1-scores, correlations with recall values are usually substantially lower. For tone-based features, we observe the converse (with one exception).

The reason for these divergences is most likely the stress and tone prediction rate of the decision tree; since the incidence of stress labels in the C4.5 training data is lower than in the human gold standard data, about 22% fewer stress labels (relative) are predicted than expected, which improves precision at the expense of recall. For tone, the picture is less clear, since although the C4.5 training data contains a larger percentage of tone labels than the human gold standard data, this does not result in a larger prediction rate by the decision tree classifier; in fact, the number of predicted tones is about 10% smaller (relative) than expected from the gold standard. Of course, the non-native speakers' prosodic characteristics are likely to diverge quite a bit from the native speaker's norm in the first place.

Table 6. Correlations between 12 prosodic features for 741 non-native read-aloud responses and human rater scores.[6]

#	Feature description	Correlation
1	Mean distance between stressed syllables (in syllables)	- 0.298
2	Mean deviation of 1	- 0.376
3	Mean distance between stressed syllables (in seconds)	- 0.040
4	Mean deviation of 3	- 0.158
5	Mean distance between syllables that bear tones (in syllables)	- 0.021
6	Mean deviation of 5	0.012
7	Mean distance between syllables that bear tones (in seconds)	0.054
8	Mean deviation of 7	0.007
9	Feature 4 / Feature 3 (normalized mean deviation of stress dist.)	- 0.195
10	Feature 8 / Feature 7 (normalized mean deviation of tone dist.)	0.112
11	Relative frequency of stressed syllables (in percent)	0.367
12	Relative frequency of syllables with boundary tone (in percent)	0.206

[6] "Mean deviation: is the average of all absolute differences between the mean of a data set and each of its elements.

In terms of our second experiment, where we computed 12 rhythm and frequency features derived directly from automatically predicted stress and tone labels, we found that the feature indicating the mean deviation of time intervals between stressed syllables had the highest absolute correlation with human rater scores (r=-0.38). This feature has high values when the time distance between stressed syllables is unevenly paced, i.e., fairly irregular. In other words, more proficient speakers are expected to produce stressed syllables on a more regular pace than lower proficient speakers do. Since more proficient speakers have higher human scores but lower feature values (more evenly paced stressed syllables), the feature correlation is negative.

Most correlations related to automated tone predictions were fairly low, due to the relatively high error rate of the tone prediction classifier. The two features representing the relative frequency of stress and tone events have somewhat reasonable correlations (r=0.37 for stress and r=0.21 for tone). They probably indicate that less proficient speakers tend to have a more flat intonation and hence produce relatively fewer syllables that bear stress or tone, as compared with their more proficient peers.

Comparing these two experiments, we note that the features that involve syllable-by syllable-comparison to expert-labeled stress and tone information outperform those that indicate overall patterns. This is not surprising as the former type provides a more fine-grained evaluation of the stress and intonation of read speech. The features that capture the deviation in overall stress patterns had stronger relationships with human scores than those related to intonation patterns, probably due to the more accurate prediction of stress and more frequent occurrence of stressed points in the data.

VI. Conclusion

This paper reports on the development and evaluation of two types of prosodic features computed using automatically labeled stress and intonation information. Both features have not previously been used for the purpose of automatically scoring non-native read speech.

The first type compares machine-annotated stress and intonation points to those annotated by expert test developers and shows the deviation in speakers' stress and intonation patterns from expected norms. The average correlations for features based on stress labels with human proficiency scores is r=0.42, and for one test item, r=0.54 is obtained. For tone-based features, the average correlation is considerably lower with r=0.31.

The second type of feature is derived based on automatically predicted stress and tone information only, and captures descriptively overall intonation and rhythmic patterns of non-native read speech. The highest absolute correlations for this feature class are r=-0.38 for the mean deviation of time intervals between stressed syllables and r=0.37 for the relative frequency of stressed syllables in a read passage. Again, features derived from tone labels exhibit lower correlations; the highest correlation is observed for the relative frequency of tones in a read passage with r=0.21.

The results show promise for using these prosodic features along with pronunciation and fluency features that have already been developed in improving the prediction of human evaluations of read speech.

In this study, we use stress and tone classifiers trained on spontaneous speech. However, the prosodic characteristics may be somewhat distinct for read speech; consequently, we plan to update the classifiers using training data from read speech and hope to improve their performance in stress and tone prediction for read speech data.

Acknowledgments

The authors wish to thank the prosodic annotators for their work, and further would like to express thanks to our colleagues Keelan Evanini, Joel Tetreault and Su-Youn Yoon, as well as to the anonymous reviewers for their valuable feedback on earlier versions of this paper.

References

[1] J. Mostow, S. Roth, A. Hauptmann, and M. Kane, "A prototype reading coach that listens," *Proceedings of the twelfth national conference on artificial intelligence (AAAI),* 1994.

[2] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: the role of multiple information sources.," *Proceedings of IEEE international workshop on multimedia signal processing,* Greece., 2007.

[3] H. Franco, H. Bratt, R. Rossier, V. R. Gadde, E. Shriberg, V. Abrash, and K. Precoda, "EduSpeak: a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications.," *Language Testing*, vol. 27, 2010.

[4] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *Journal of the Acoustical Society of America*, vol. 111, pp. 2862-2873, June 2002.

[5] J. Liscombe, "Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency," Ph.D. thesis, New York, NY: Columbia University, 2007.

[6] T. Y. Jang, "Speech rhythm metrics for automatic scoring of english speech by Korean EFL learners," *Malsori (Speech Sounds) The Korean Society of Phonetic Sciences and Speech Technology*, vol. 66, pp. 41–59, 2008.

[7] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental English through parroting," *Proc. of Speech Prosody*, 2010.

[8] (2011) Praat: Doing phonetics by computer [Online]. Available: http://www.fon.hum.uva.nl/praat/

[9] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in Prosodic Typology -- The Phonology of Intonation and Phrasing, S. A. Jun, Ed. Oxford, UK: Oxford University Press, 2005.

[10] Linguistic Data Consortium, "HUB-4 Broadcast News corpus (English)," 1997.

[11] L. Chen, K. Zechner, and X. Xi, "Improved Pronunciation Features for Construct-Driven Assessment of Non-Native Spontaneous Speech. ," *Proceedings of the NAACL-HLT-2009 Conference*, Boulder, CO, 2009.

[12] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1992.

[13] (2011) NIST: Rich Transcription 2003 Evaluation [Online] Available: http://www.itl.nist.gov/iad/mig/tests/rt/2003-fall/index.html