Predicting Word Accuracy for the Automatic Speech Recognition of Non-Native Speech

Su-Youn Yoon, Lei Chen, Klaus Zechner

Educational Testing Service, Princeton, NJ, USA

{syoon,lchen,kzechner}@ets.org

Abstract

We have developed an automated method that predicts the word accuracy of a speech recognition system for non-native speech, in the context of speaking proficiency scoring. A model was trained using features based on speech recognizer scores, function word distributions, prosody, background noise, and speaking fluency.

Since the method was implemented for non-native speech, fluency features, which have been used for non-native speakers' proficiency scoring, were implemented along with several feature groups used from past research. The fluency features showed promising performance by themselves, and improved the overall performance in tandem with other more traditional features.

A model using stepwise regression achieved a correlation with word accuracy rates of 0.76, compared to a baseline of 0.63 using only confidence scores. A binary classifier for placing utterances in high-or low-word accuracy bins achieved an accuracy of 84%, compared to a majority class baseline of 64%.

Index Terms: speech recognition, word accuracy rate, non-native speakers' speech

1. Introduction

In this paper, we develop a word accuracy rate (WAR)[1] prediction method as a supplementary module of the automated speech proficiency scoring system SpeechRater^{TM}. SpeechRater is the automated scoring system of Educational Testing Service (ETS) [1]. SpeechRater scores the proficiency of non-native speakers' spontaneous speech as part of the TOEFL® Practice Online (TPO), a low-stakes practice test product.

Recognition of non-native speakers' speech is, by itself, a very challenging task and is even more difficult in our case, due to the large diversity of speaker proficiencies and native language backgrounds. While features related to fluency, such as speaking rate, can be estimated fairly reliably even with moderate word accuracies, this is not the case for features related to the higher-level aspects of proficiency, such as vocabulary diversity, grammatical complexity and correctness, or topical coherence. In all of these instances, features rely heavily on correctly recognized words and so utterances that have word accuracies that are too low may have to be identified for special processing. Therefore, our automated method for estimating the WAR can be a useful tool to determine whether features representing higher-level aspects of speaking proficiency are reliable or not.

The overall architecture of our method is as follows: for a given speech response, SpeechRater performs speech recognition, yielding a word hypothesis and time stamps, and computes basic prosodic features (pitch and power). Next, it computes fluency features for automated proficiency scoring. Finally, the WAR prediction method estimates the WAR, using output from the automated speech recognition (ASR) system, fluency features, and a number of additional features described below.

This paper will proceed as follows: we will review previous studies (Section 2), present the automated WAR prediction method (Section 3), and then report the experimental setup (Section 4). Next, the results will be presented (Section 5) and compared with the previous studies in depth (Section 6).

2. Previous work

WAR prediction systems have been widely used in ASR research. In particular, methods based on features from the speech recognizer, such as confidence scores and N-best candidates [2,3], have rendered promising performances. For example, the word error detection system in [3] rejected 65.7% of the errors correctly while falsely rejecting 5.1% of the correct words.

It has been used as an important module for dialogue management systems [4–6]. Here, the WAR prediction method allows for more effective communication between machines and humans; it helps to make the decision whether to generate responses based on the current word hypotheses or reject them and request the same information again.

Litman et al. [4] found a significant difference in prosody between speaker turns with and without recognition errors. The speech recognizer tends to make more recognition errors in very loud and fast speech. In order to identify these potential error regions, the authors used features such as amplitude, pitch, and speaking rate. These prosodic features outperformed the speech recognizer-based features in their train reservation corpus. The combination of the two feature groups achieved 93.5% accuracy in misrecognized turn detection.

Discourse and pragmatic features were also used in previous research. Walker et al. [5] developed a system based on pragmatic features (for example, the dialogue move type), discourse history, speech recognizer-based features and prosodic features, and it achieved 86% accuracy in error detection.

The current study can be distinguished from the previous studies in the following points. First of all, special features

[1] In this study, the balanced word accuracy rate was used to measure speech recognition performance. WAR is the mean between the reference-based and the hypothesis-based word accuracy. WAR is calculated as follows:

$$\mathrm{WAR} = \frac{1}{2}\left\{\frac{\mathrm{C}}{(\mathrm{C+S+D})} + \frac{\mathrm{C}}{(\mathrm{C+S+I})}\right\}.$$

where C = number of the correct words in the word hypothesis, S = number of the substituted words in the word hypothesis, D = number of the deleted words in the word hypothesis, and I = number of the inserted words in the word hypothesis.

26 – 30 September 2010, Makuhari, Chiba, Japan

were implemented to model non-native speech since the method was developed for non-native speech. Non-native speakers tend to produce errors and disfluencies more frequently than native speakers, and the ASR system tends to make more errors in these non-fluent regions of speech. In order to identify these non-native characteristics, fluency features related to speaking rate and disfluency, which achieved good performance in the estimation of the non-native speakers' speaking proficiency [1, 7, 8], were implemented. Secondly, in contrast to previous studies which made a categorical decision as to whether a specific word or utterance contained a recognition error or not, we estimate the WAR of a response (file) with a continuous real value. Finally, various normalization methods were applied for the features commonly used in previous studies.

3. Method

3.1. Task

The method estimates the WAR using speech recognizer scores, function word distributions, noise-related features, prosodic features (such as F_0 and power), and fluency features.

3.2. Features for WAR prediction

Five different groups of features were selected and calculated automatically. A list of features is provided in Table 1.

The category and the name of the feature are presented in the first and second columns, respectively. New features proposed by this study (new features) are presented in bold text, while features similar to those used in previous studies such as [4–6] (basic features) are presented in plain text.

The third column shows the correlation with WAR for the training data described in 4.1. Some of the new features achieved stronger correlations than the basic features. For instance, the mean histogram-probability feature showed the highest correlation with WAR (r=0.64), and it was stronger than the correlation of the mean raw confidence score which achieved the highest correlation among the basic features (r=0.62).

Finally, the fourth column presents whether the feature was actually used in the final method or not. A detailed explanation will be presented in Section 4.3.

3.2.1. Speech recognizer features

The speech recognizer-based features were computed for each individual word, and the mean and standard deviation of the scores were computed. In general, the confidence score is a promising feature in WAR prediction, but it has a skewed distribution towards high values, and the overall high scores are not always reliable. Due to this bias, the probability that the word may be correct given the confidence score has been used instead of the raw score [9]. In this study, the raw score was mapped to the probability (histogram-probability) using a histogram. The confidence scores of hypothesized words in the training data were classified into 10 bins, and a histogram was constructed. The histogram-probability $P(q_j)$ for the raw score q_j was computed by:

$$P(q_j) = \frac{N(S_j, label=1)}{N(S_j, label=-1)+N(S_j, label=1)}$$
where $q_j \in S_j$ and, $N(S_j, label = 1)$ is the number of positive examples in score bin S_j (1)

Table 1: List of all features for the WAR prediction method (Features in bold text are new features proposed by this study)

Feature category	Feature name	Corr.	Model
Speech recognizer	Mean Acoustic model score[a]	-0.34**	8
	Mean Language model score[b]	-0.30**	No
	Mean raw confidence scores	0.62**	No
	STD.[c] of raw confidence scores	-0.38**	No
	Mean histogram-probability	0.64**	1
	Proportion of low confidence scores	-0.50**	6
	Confidence score per second	0.58**	7
Function word distribution	**Frequency**	-0.22**	10
	Proportion	-0.23**	12
Prosody	Mean power	0.08	No
	Max power	-0.09*	No
	Min power	0.02	No
	STD. of pitch normalized by speaker	0.13**	5
	Max pitch	0.08*	No
	Min pitch	-0.05	No
Background noise	Signal to noise ratio	0.05	No
	Mean noise level	-0.02	No
	Peak speech level	0.10	No
Fluency	Words per second	0.37**	2
	Mean long silence duration	0.11**	3
	STD. of long silence duration	0.10*	11
	Silences per second	0.11**	9
	Frequency of disfluency[d]	-0.12**	4

* Correlation is significant at the 0.05 level
** Correlation is significant at the 0.01 level
[a] Sum of the log probabilities of the reference acoustic model normalized by number of phones
[b] Sum of the log probabilities of the reference language model normalized by number of words
[c] Standard deviation
[d] Number of disfluencies such as pauses, fillers, or repetitions in the word hypotheses

3.2.2. Function word distributions

Out-of-vocabulary words tend to be recognized as a sequence of short function words. Therefore, if the speech contains out-of-vocabulary words, the word hypotheses may include more function words. In order to capture this characteristic, the frequency and proportion of function words in the word hypotheses were calculated.

Function word distribution features were computed using the SMART function word list from [10]. From the word hypothesis, the frequency and the proportion of function words were computed.

3.2.3. Prosodic features

Prosodic features (pitch and power) were implemented to identify very loud or abnormal speech. Pitch and power were obtained using a pitch and power extraction module in the speech recognizer.

3.2.4. Noise related features

The speech recognizer tends to make more recognition errors in a noisy environment than in a quiet environment. The signal to noise ratio (SNR), mean noise level, and peak speech level were computed using the NIST audio quality assurance (SPQA) package [11].

3.2.5. Fluency features

Fluency features were implemented to identify non-fluent speech. Speaking rate and pause-related features, such as the duration and proportion of total speech, were used in this study.

Fluency features were calculated from the word hypotheses yielding the number of words, the duration of responses, and the number and duration of silences.

3.3. Model

Since the predicted value is a real number, a multiple regression model was used for model building. The regression model was built using a selected set of basic and new features. The criterion for selection was a stepwise linear regression analysis using the training data.

4. Experiment

4.1. Data

In this study, data from the TOEFL®Practice Online (TPO) were used for both training and testing. The TPO assessment consists of 6 items to which speakers are prompted to provide responses between 45 and 60 seconds per item. The scoring scale is discrete from 1 to 4, where '4' indicates high and '1' low-speaking proficiency. Additionally, a score of '0' is used to indicate a non-response and a score of 'TD' to indicate technical difficulties such as static noise that prevented the response from being scored.

Responses which were scored as '0' or 'TD' were excluded from the study, and a total of 1040 responses were used for the development and testing of the automated WAR prediction method. 645 responses were used for training, while 395 responses were used for testing. There is no speaker overlap among the training and test data.

The speakers varied widely in their English proficiency levels; the data included fluent speakers, intermediate speakers, and also a few low-proficiency speakers. Detailed information about the rating process and proportion at each proficiency level for the TPO data can be found in [1].

4.2. Speech recognizer

The acoustic model of a gender-independent HMM recognizer was trained on approximately 30 hours of non-native speech (TPO data), and a language model was trained using both native data (1997 Broadcast news data) and non-native data (TPO data). The WAR on the test data was around 50%. In addition to word hypotheses, the speech recognizer generates a confidence score for each word from 0.0 to 1.0.

4.3. Feature selection and model building

First, all features in Table 1 were computed automatically for each response.

Many features had a significant correlation with WAR, but they were also strongly correlated with other features. In order to investigate which combination of features can improve the regression model, a stepwise linear regression analysis was performed using the training data using the SPSS statistical analysis program. The order selected in the model is presented in the fourth column of Table 1. If the feature was not used in the regression model, the column is labeled 'No'.

A total of 12 features were selected. Among these 12 features, 2 features were from the basic group and 10 features were from the new group. The best predictor was histogram-probability, followed by words per second. The new features in the fluency group and function word distribution did not achieve strong correlations, but they improved the performance of the regression model.

Finally, a linear regression model was trained using the 12 features. The WEKA machine learning toolkit [12] was used for training and testing the model.

5. Results

5.1. Regression

Table 2 shows the performance of the regression models on the test data.

Table 2: Performance of regression models using different sets of features

Features	RMS Error[a]	RRS Error[b]	Correlation[c]
Baseline	0.11	74.8	0.63
Total	0.09	64.3	0.76

[a] Root mean-squared error
[b] Root relative-squared error
[c] Pearson correlation coefficient

The baseline model was trained using only the mean raw confidence scores, since it has been commonly used in WAR estimation, and was also used as a baseline in [4]. The total model improved significantly over the baseline; the correlation increased from 0.63 to 0.76.

In order to measure the impact of the features on WAR prediction, the original complete set of features were classified into five groups (as shown in Table 1) and a regression model was trained separately for each group. The best features were the speech recognizer-related features with a correlation of 0.65 with WAR. This was followed by fluency ($r = 0.56$), function word distribution ($r = 0.18$). Background noise and prosody features did not show significant correlations with WAR.

5.2. Binary classification

In order to see whether the regression model could be used in a scenario where binary choices need to be made, for example, to allow the computation of higher-level features for an utterance or not (as discussed earlier), we conducted a small experiment where we classified utterances in the test set to 'low' or 'high' WAR, depending on the output of the regression model. As a threshold we chose WAR$= 0.6$, which puts 64% of the test responses into the 'low' bin and the rest into the 'high' bin.

The classification accuracy of this binary task was 84%, a 20% absolute improvement over the majority class baseline.

6. Discussion

Many features used in this study were similar to the features used in [4–6]. However, the performance of these features showed different tendencies with our data.

The contribution of the prosody features was weak for this study. The difference in speech style may be relevant to this result. In dialogue applications, the speaker tends to speak louder or raise his/her pitch when the recognizer misrecognizes speech. This may not happen in our data, since the data is a monologue. These differences may weaken the relationship between the prosody features and WAR.

Similarly, background noise features did not correlate well with WAR in our data. One possibility to consider is the noise type: Fish et al. [13] showed that the influence on WAR differs according to the noise type. The implementation of the appropriate features for different types of noise may enhance the relationship between noise-related features and WAR. Additionally, there is not a lot of noise present in most TPO responses; approximately 90% of the responses are classified by human raters as having high or very high audio quality.

In this study, the speaking rate was the second best predictor. It had a significant positive correlation with WAR; that is, if a speaker speaks fast, the accuracy of recognition increases. This result seems to be inconsistent with [4, 14]; they found that misrecognized word groups are significantly faster than correctly recognized word groups.

A close comparison, however, can resolve the inconsistency between these results. In [14], the recognition errors were influenced by the speaking rate only when it was very high. The speaking rate did not influence WAR when speakers spoke at a normal speed[2]. In our data, no speakers spoke faster than a normal speed since they are all non-native speakers. This explains why a negative correlation between WAR and the speaking rate is not found in this study.

The relationship between the speaking rate and non-native speakers' proficiency may explain the positive correlation between the speaking rate and WAR. The speaking rate is a prominent factor for estimating non-native speakers' proficiency levels, and it has been widely deployed in automated speech proficiency scoring systems [1,8]. Speakers who speak faster tend to be more fluent than speakers who speak more slowly. As speakers become more fluent, they make fewer errors, and WAR may increase. Thus, the positive correlation between the proficiency and speaking rate may result in the positive correlation between WAR and the speaking rate.

7. Conclusions

In this study, we presented a WAR prediction method for non-native speakers' speech. A regression model was trained based on speech recognizer scores, function word distributions, prosody, background noise, and fluency features.

The method is intended to function as a supplementary module of an automated speech proficiency scoring system. Thus, the method implemented fluency features which were specialized for second language (L2) learners' speech, and fluency features have proven to be very effective in estimating non-native speakers' speech proficiency. The strong relationship between non-native speakers' proficiency and the accuracy of the speech recognition system also contributes to the predictive power of the fluency features; they were the second-best predictor of WAR.

[2]WAR decreased sharply when the speaking rate was faster than 12 phones per second.

8. Acknowledgements

We thank Keelan Evanini and Derrick Higgins of Educational Testing Service for their comments and suggestions.

9. References

[1] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, May 2009.

[2] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition," in *Proceedings of the European Conference on Speech Communication and Technology*, 1997, pp. 815–818.

[3] T. J. Hazen, T. Burianek, J. Polifroni, and S. Seneff, "Recognition confidence scoring and its use in speech understanding systems," *Computer Speech and Language*, pp. 49–67, 2002.

[4] D. Litman, J. Hirshberg, and M. Swerts, "Predicting automatic speech recognition performance using prsodic cues," in *Proceedings of the 6th International Conference of Spoken Language Processing*, 2000.

[5] M. Walker, J. Wright, and I. Langkilde, "Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system," in *Proceedings of the 17th International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 1111–1118.

[6] M. Gabsdil, "Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems," in *Proceedings of ACL*, 2004, pp. 344–351.

[7] P. Lennon, "Investigating fluency in EFL: A quantitative approach," *Language Learning*, vol. 40, no. 3, pp. 387–417, 1990.

[8] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *the Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[9] C. Neti, S. Roukos, and E. Eide, "Word-based confidence measures as a guide for stack search in speech recognition," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE Computer Society, 1997, pp. 883–886.

[10] Salton, G., Ed., *The SMART Retrieval System:Experiments in Automatic Document Processing*. Prentice-Hall, 1971.

[11] NIST, "The nist speech quality assurance (spqa) package version 2.3," from http://www.nist.gov/speech/tools/index.htm, Retrieved 2009.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," in *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[13] R. Fish, Q. Hu, and S. Boykin, "Using audio quality to predict word error rate in an automatic speech recognition system," Unpublished 2006 manuscript from The MITRE Corporation.

[14] S. Goldwater, D. Jurafsky, and C. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, pp. 181–200, 2010.