Exploring Content Features for Automated Speech Scoring

Shasha Xie, Keelan Evanini, Klaus Zechner
Educational Testing Service (ETS)
Princeton, NJ 08541, USA
{sxie,kevanini,kzechner}@ets.org

Abstract

Most previous research on automated speech scoring has focused on restricted, predictable speech. For automated scoring of unrestricted spontaneous speech, speech proficiency has been evaluated primarily on aspects of pronunciation, fluency, vocabulary and language usage but not on aspects of content and topicality. In this paper, we explore features representing the accuracy of the content of a spoken response. Content features are generated using three similarity measures, including a lexical matching method (Vector Space Model) and two semantic similarity measures (Latent Semantic Analysis and Pointwise Mutual Information). All of the features exhibit moderately high correlations with human proficiency scores on human speech transcriptions. The correlations decrease somewhat due to recognition errors when evaluated on the output of an automatic speech recognition system; however, the additional use of word confidence scores can achieve correlations at a similar level as for human transcriptions.

1 Introduction

Automated assessment of a non-native speaker's proficiency in a given language is an attractive application of automatic speech recognition (ASR) and natural language processing (NLP) technology; the technology can be used by language learners for individual practice and by assessment providers to reduce the cost of human scoring. While much research has been done about the scoring of restricted speech, such as reading aloud or repeating sentences verbatim (Cucchiarini et al., 1997; Bernstein et al., 2000; Cucchiarini et al., 2000; Witt and Young, 2000; Franco et al., 2000; Bernstein et al., 2010b), much less has been done about the scoring of spontaneous speech. For automated scoring of unrestricted, spontaneous speech, most automated systems have estimated the non-native speakers' speaking proficiency primarily based on low-level speaking-related features, such as pronunciation, intonation, rhythm, rate of speech, and fluency (Cucchiarini et al., 2002; Zechner et al., 2007; Chen et al., 2009; Chen and Zechner, 2011a), although a few recent studies have explored features based on vocabulary and grammatical complexity (Zechner et al., 2007; Bernstein et al., 2010a; Bernstein et al., 2010b; Chen and Zechner, 2011b).

To date, little work has been conducted on automatically assessing the relatively higher-level aspects of spontaneous speech, such as the content and topicality, the structure, and the discourse information. Automated assessment of these aspects of a non-native speaker's speech is very challenging for a number of reasons, such as the short length of typical responses (approximately 100 words for a typical 1 minute response, compared to over 300 words in a typical essay/written response), the spontaneous nature of the speech, and the presence of disfluencies and possible grammatical errors. Moreover, the assessment system needs text transcripts of the speech to evaluate the high level aspects, and these are normally obtained from ASR systems. The recognition accuracy of state-of-the-art ASR systems on non-native spontaneous speech is still relatively low, which will sequentially impact the re-

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 103–111,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

liability and accuracy of automatic scoring systems using these noisy transcripts. However, despite these difficulties, it is necessary for an automated assessment system to address the high level information of a spoken response in order to fully cover all aspects that are considered by human raters. Thus, in this paper we focus on exploring features to represent the high-level aspect of speech mainly on the accuracy of the content.

As a starting point, we consider approaches that have been used for the automated assessment of content in essays. However, due to the qualitative differences between written essays and spontaneous speech, the techniques developed for written texts may not perform as well on spoken responses. Still, as a baseline, we will evaluate the content features used for essay scoring on spontaneous speech. In addition to a straightforward lexical Vector Space Model (VSM), we investigate approaches using two other similarity measures, Latent Semantic Analysis (LSA) and Pointwise Mutual Information (PMI), in order to represent the semantic-level proficiency of a speaker. All of the content features are analyzed using both human transcripts and speech recognizer output, so we can have a better understanding of the impact of ASR errors on the performance of the features. As expected, the results show that the performance on ASR output is lower than when human transcripts are used. Therefore, we propose improved content features that take into account ASR confidence scores to emphasize responses whose estimated word accuracy is comparatively higher than others. These improved features can obtain similar performance when compared to the results using human transcripts.

This paper is organized as follows. In the next section we introduce previous research on automated assessment of content in essays and spoken responses. The content features we generated and the model we used to build the final speaking scores are described in Sections 3 and Section 4, respectively. In Section 5 we show the performance of all our proposed features. Finally, we conclude our work and discuss potential future work in Section 6.

2 Related Work

Most previous research on assessment of non-native speech has focused on restricted, predictable speech; see, for example, the collection of articles in (Eskenazi et al., 2009). When assessing spontaneous speech, due to relatively high word error rates of current state-of-the-art ASR systems, predominantly features related to low-level information have been used, such as features related to fluency, pronunciation or prosody (Zechner et al., 2009).

For scoring of written language (automated essay scoring), on the other hand, several features related to the high level aspects have been used previously, such as the content and the discourse information. In one approach, the lexical content of an essay was evaluated by using a VSM to compare the words contained in each essay to the words found in a sample of essays from each score category (Attali and Burstein, 2006). In addition, this system also used an organization feature measuring the difference between the ideal structure of an essay and the actual discourse elements found in the essay. The features designed for measuring the overall organization of an essay assumed a writing strategy that included an introductory paragraph, at least a three-paragraph body with each paragraph in the body consisting of a pair of main point, supporting idea elements, and a concluding paragraph. In another approach, the content of written essays were evaluated using LSA by comparing the test essays with essays of known quality in regard of their degree of conceptual relevance and the amount of relevant content (Foltz et al., 1999).

There has been less work measuring spoken responses in terms of the higher level aspects. In (Zechner and Xi, 2008), the authors used a content feature together with other features related to vocabulary, pronunciation and fluency to build an automated scoring system for spontaneous high-entropy responses. This content feature was the cosine word vector product between a test response and the training responses which have the highest human score. The experimental results showed that this feature did not provide any further contribution above a baseline of only using non-content features, and for some tasks the system performance was even slightly worse after including this feature. However, we think the observations about the content features used in this paper were not reliable for the following two reasons: the number of training responses was limited (1000 responses), and the ASR system had a relatively high Word Error Rate (39%).

In this paper, we provide further analysis on the performance of several types of content features. Additionally, we used a larger amount of training data and a better ASR system in an attempt to extract more meaningful and accurate content features.

3 Automatic Content Scoring

In automatic essay scoring systems, the content of an essay is typically evaluated by comparing the words it contains to the words found in a sample of essays from each score category (1-4 in our experiments), where the scores are assigned by trained human raters. The basic idea is that good essays will resemble each other in their word choice, as will poor essays. We follow this basic idea when extracting content features for spoken responses.

3.1 Scoring Features

For each test spoken response, we calculate its similarity scores to the sample responses from each score category. These scores indicate the degree of similarity between the words used in the test response and the words used in responses from different score points. Using these similarity scores, 3 content features are generated in this paper:

- Sim_{max}: the score point which has the highest similarity score between test response and score vector
- Sim_4: the similarity score to the responses with the highest score category (4 in our experiments).
- Sim_{cmb}: the linear combination of the similarity scores to each score category.

$$\sum_{i=1}^{4} w_i * Sim_i \tag{1}$$

where w_i is scaled to [-1, 1] to imply its positive or negative impact.

3.2 Similarity Measures

There are many ways to calculate the similarity between responses. A simple and commonly used method is the Vector Space Model, which is also used in automated essay scoring systems. Under this approach, all the responses are converted to vectors, whose elements are weighted using TF*IDF (term frequency, inverse document frequency). Then, the cosine similarity score between vectors can be used to estimate the similarity between the responses the vectors originally represent.

Other than this lexical matching method, we also try two additional similarity measures to better capture the semantic level information: Latent Semantic Analysis (Landauer et al., 1998) and a corpus-based semantic similarity measure based on pointwise mutual information (Mihalcea et al., 2006). LSA has been widely used for computing document similarity and other information retrieval tasks. Under this approach, Singular Value Decomposition (SVD) is used to analyze the statistical relationship between a set of documents and the words they contain. A $m*n$ word-document matrix X is first built, in which each element X_{ij} represents the weighted term frequency of word i in document j. The matrix is decomposed into a product of three matrices as follows:

$$X = U\Sigma V^T \tag{2}$$

where U is an $m \times m$ matrix of left-singular vectors, Σ is an $m \times n$ diagonal matrix of singular values, and V is the $n \times n$ matrix of right-singular vectors.

The top ranked k singular values in Σ are kept, and the left is set to be 0. So Σ is reformulated as Σ_k. The original matrix X is recalculated accordingly,

$$X_k = U\Sigma_k V^T \tag{3}$$

This new matrix X_k can be considered as a smoothed or compressed version of the original matrix. LSA measures the similarity of two documents by calculating the cosine between the corresponding compressed column vectors.

PMI was introduced to calculate the semantic similarity between words in (Turney, 2001). It is based on the word co-occurrence on a large corpus. Given two words, their PMI is computed using:

$$PMI(w_1, w_2) = log_2 \frac{p(w_1 \& w_2)}{p(w_1) * p(w_2)} \tag{4}$$

This indicates the statistical dependency between w_1 and w_2, and can be used as a measure of the semantic similarity of two words.

Given the word-to-word similarity, we can calculate the similarity between two documents using the following function,

$$sim(D_1, D_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{D_1\}}(maxSim(w, D_2) * idf(w))}{\sum_{w\in\{D_1\}} idf(w)} + \frac{\sum_{w\in\{D_2\}}(maxSim(w, D_1) * idf(w))}{\sum_{w\in\{D_2\}} idf(w))}\right) \quad (5)$$

$$maxSim(w, D_i) = max_{w_j \in \{D_i\}} PMI(w, w_j) \quad (6)$$

For each word w in document D_1, we find the word in document D_2 which has the highest similarity to w. Similarly, for each word in D_2, we identify the most similar words in D_1. The similarity score between two documents is then calculated by combining the similarity of the words they contain, weighted by their word specificity (i.e., IDF values).

In this paper, we use these three similarity measures to calculate the similarity between the test response and the training responses for each score category. Using the VSM method, we convert all the training responses in one score category into one big vector, and for a given test response we calculate its cosine similarity to this vector as its similarity to that corresponding score point vector. For the other similarity measures, we calculate the test response's similarity to each of the training responses in one score category, and report the average score as its similarity to this score point. We also tried using this average similarity score for the VSM method, but our experimental results showed that this average score obtained lower performance than using one big vector generated from all the training samples due to data sparsity. After the similarity scores to each of the four score categories are computed, the content features introduced in Section 3.1 are then extracted and are used to evaluate the speaking proficiency of the speaker.

4 System Architecture

This section describes the architecture of our automated content scoring system, which is shown in Figure 1. First, the test taker's voice is recorded, and sent to the automatic speech recognition system. Second, the feature computation module takes the output hypotheses from the speech recognizer and generates the content features. The last component considers all the scoring features, and produces the final score for each spoken response.

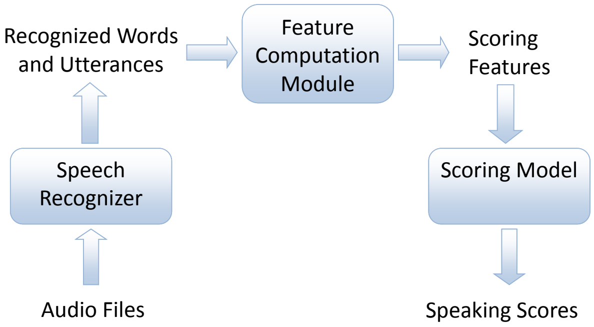

Figure 1: Architecture of the automated content scoring system.

While we are using human transcripts of spoken responses as a baseline in this paper, we want to note that in an operational system as depicted in this figure, the scoring features are computed and extracted using the hypotheses from the ASR system, which exhibits a relatively high word error rate. These recognition errors will sequentially impact the process of calculating the similarity and computing the content scores, and decrease the performance of the final speaking scores. In order to improve the system performance in this ASR condition, we explore the use of word confidence scores from the ASR system during feature generation. In particular, the similarity scores between the test response and each score category are weighted using the recognition confidence score of the response, so that the scores can also contain information related to its acoustic accuracy. The confidence score for one response is the average value of all the confidence scores for each word contained in the response. In Section 5, we will evaluate the performance of our proposed content features using both human transcripts and ASR outputs, as well as the enhanced content features using ASR confidence scores.

5 Experimental Results

5.1 Data

The data we use for our experiments are from the Test of English as a Foreign Language® internet-based test (TOEFL iBT) in which test takers respond to several stimuli using spontaneous speech. This data set contains 24 topics, of which 8 are opinion-based tasks, and 16 are contextual-based tasks. The opinion-based tasks ask the test takers to provide information or opinions on familiar topics based on their personal experience or background knowledge. The purpose of these tasks is to measure the speaking ability of examinees independent of their ability to read or listen to English language. The contextual-based tasks engage reading, listening and speaking skills in combination to mimic the kinds of communication expected of students in campus-based situations and in academic courses. Test takers read and/or listen to some stimulus materials and then respond to a question based on them. For each of the tasks, after task stimulus materials and/or test questions are delivered, the examinees are allowed a short time to consider their response and then provide their responses in a spontaneous manner within either 45 seconds (for the opinion-based tasks) or 60 seconds (for the contextual-based tasks).

For each topic, we randomly select 1800 responses for training, and 200 responses as development set for parameter tuning. Our evaluation data contains 1500 responses from the same English proficiency test, which contain the same 24 topics. All of these data are scored on a 0-4 scale by expert human raters. In our automated scoring system, we use a filtering model to identify responses which should have a score of 0, such as responses with a technical difficulty (e.g., equipment problems, high ambient noise), responses containing uncooperative behavior from the speakers (e.g., non-English speech, whispered speech). So in this paper we only focused on the responses with scores of 1-4. Statistics for this data set are shown in Table 1. As the table shows, the score distributions are similar across the training, development, and evaluation data sets.

5.2 Speech recognizer

We use an ASR system containing a cross-word triphone acoustic model trained on approximately 800 hours of spoken responses from the same English proficiency test mentioned above and a language model trained on the corresponding transcripts, which contain a total of over 5 million words. The Word Error Rate (WER) of this system on the evaluation data set is 33%.

5.3 Evaluation metric

To measure the quality of the developed features, we employ a widely used metric, the Pearson correlation coefficient (r). In our experiments, we use the value of the Pearson correlation between the feature values and the human proficiency scores for each spoken response.

5.4 Feature performance on transcripts

In Section 3.1, we introduced three features derived from the similarity between the test responses and the training responses for each score point. We first build the training samples for each topic, and then compare the test responses with their corresponding models. Three similarity measures are used for calculating the similarity scores, VSM, LSA, and the PMI-based method. In order to avoid the impact of recognition errors, we first evaluate these similarity methods and content features using the human transcripts. The Pearson correlation coefficients on the evaluation data set for this experiment are shown in Table 2. The parameters used during model building, such as the weights for each score category in the feature Sim_{cmb} and the number of topics k in LSA, are all tuned on the development set, and applied directly on the evaluation set.

The correlations show that even the simple vector space model can obtain a good correlation of 0.48 with the human rater scores. The feature Sim_{cmb} performs the best across almost all the test setups, since it combines the information from all score categories. The PMI-based features outperform the other two similarity methods when evaluated both on all responses or only on the contextual-based topics. We also observe that the correlations on *contextual-based* tasks are much higher than on *opinion-based* tasks. The reason for this is that

Table 1: Summary statistics of training, development and evaluation data set.

Data sets	*Responses*	*Speakers*	*score_avg*	*score_sd*	*Score distribution (percentage %)*			
					1	2	3	4
Train	43200	8000	2.63	0.79	1750 (4.1)	15128 (35.0)	20828 (48.2)	4837 (11.2)
Dev	4800	3760	2.61	0.79	215 (4.5)	1719 (35.8)	2295 (47.8)	499 (10.4)
Eval	1500	250	2.57	0.81	95 (6.3)	549 (36.6)	685 (45.7)	152 (10.1)

Table 2: Pearson correlations of the content features using human transcripts.

	VSM			*LSA*			*PMI*		
	Sim_{max}	Sim_4	Sim_{cmb}	Sim_{max}	Sim_4	Sim_{cmb}	Sim_{max}	Sim_4	Sim_{cmb}
ALL	0.46	0.32	**0.48**	0.32	0.38	**0.45**	0.18	0.51	**0.53**
Contextual	0.50	0.51	**0.58**	0.36	0.55	**0.57**	0.21	0.57	**0.62**
Opinion	**0.37**	0.03	0.25	**0.29**	0.14	0.22	0.06	0.42	**0.51**

the contextual-based tasks are more constrained to the materials provided with the test item, whereas the opinion-based tasks are relatively open-ended. Therefore, it is easier for the similarity measures to track the content, the topics, or the vocabulary usage of the contextual-based topics. Overall, the best correlations are obtained using the feature combining the similarity scores to each score category and the PMI-based methods to calculate the similarity. Here, the Pearson correlations are 0.53 for all responses, and 0.62 for the contextual-based tasks only.

We also investigated whether additional performance gains could be achieved by combining information from the three different content features to build a single overall content score, since the three features may measure disparate aspects of the response. The combination model we use is multiple regression, in which the score assigned to a test response is estimated as a weighted linear combination of a selected set of features. The features are the similarity values to each score category ($Sim_i, i \in \{1, 2, 3, 4\}$), calcuated using the three similairty measures. In total we have 12 content features. The regression model is also built on the development set, and tested on the evaluation set. The correlation for the final model is 0.60 on all responses, which is significantly better than the individual models (0.48 for VSM, 0.45 for LSA, and 0.53 for PMI). Compared to results reported in previous work on similar speech scoring tasks but measuring other aspects of speech, our correlation results are very competitive (Zechner and Xi, 2008; Zechner et al., 2009).

5.5 Feature Performance on ASR output

The results shown in the previous section were obtained using human transcripts of test responses, and were reported in order to demonstrate the meaningfulness of the proposed features. However, in practical automated speech scoring systems, the only available text is the output of the ASR system, which may contain a large number of recognition errors. Therefore, in this section we show the performance of the content features extracted using ASR hypotheses. Note that we still use the human transcripts of the training samples to train the models, the parameter values and the regression weights; however, we only use ASR output of the evaluation data for testing the feature performance. These correlations are shown in Table 3.

Compared to the results in Table 2, we find that the VSM and LSA methods are very robust to recognition errors, and we only observe slight correlation decreases on these features. However, the decrease for the PMI-based method is quite large. A possible reason for this is that this method is based on word-to-word similarity computed on the training data, so the mismatch between training and evaluation set likely has a great impact on the computation of the similarity scores, since we train on human transcripts, but test using ASR hypotheses. Likely for the same reason, the regression model combining all the features does not provide any further contribution to the correlation result (0.44 when evaluated

Table 3: Pearson correlations of the content features using ASR output.

	VSM			*LSA*			*PMI*		
	Sim_{max}	Sim_4	Sim_{cmb}	Sim_{max}	Sim_4	Sim_{cmb}	Sim_{max}	Sim_4	Sim_{cmb}
ALL	0.43	0.34	**0.48**	0.30	0.37	**0.43**	0.11	0.24	**0.42**
Contextual	0.49	0.53	**0.58**	0.34	0.54	**0.57**	0.16	0.31	**0.53**
Opinion	**0.30**	0.05	0.07	**0.25**	0.12	0.15	0.05	0.17	**0.27**

Table 4: Pearson correlations of the content features using ASR output with confidence scores.

	VSM			*LSA*			*PMI*		
	Sim_{max}	Sim_4	Sim_{cmb}	Sim_{max}	Sim_4	Sim_{cmb}	Sim_{max}	Sim_4	Sim_{cmb}
ALL	0.43	0.36	**0.48**	0.30	0.40	**0.45**	0.11	0.39	**0.51**
Contextual	0.49	0.55	**0.58**	0.34	0.57	**0.59**	0.16	0.46	**0.59**
Opinion	**0.30**	0.24	0.25	**0.25**	0.18	0.20	0.05	0.32	**0.40**

on all responses).

In Section 4, we proposed using ASR confidence scores during feature extraction to introduce acoustic level information and, thus, penalize responses for which the ASR output is less likely to be correct. Under this approach, all similarity scores are multiplied by the average word confidence score contained in the test response. The performance of these enhanced features is provided in Table 4. Compared to the scores in Table 3, the enhanced features perform better than the basic features that do not take the confidence scores into consideration. Using this approach, we can improve the correlation scores for most of the features, especially for the PMI-based features. These features had lower correlations because of the recognition errors, but with the confidence scores, they outperform the other features when evaluated both on all responses or only on contextual-based responses. Note that the correlations for feature Sim_{max} remains the same because the same average confidence scores for each test response is multiplied by the similarity scores to each of the score points, so the score point obtaining the highest similarity score is the same whether the confidence scores are considered or not. The correlation of the regression model also improves from 0.44 to 0.51 when the confidence scores are included. Overall, the best correlations for the individual similarity features with the confidence scores are very close to those obtained using human transcripts, as shown in Tables 2 and 4: the difference is 0.53 vs. 0.51 for all responses, and 0.62 vs. 0.59 for contextual-based tasks only.

Because all models and parameter values are trained on human transcripts, this experimental setup might not be optimal for using ASR outputs. For instance, the regression model does not outperform the results of individual features using ASR outputs, although the confidence scores help improve the overall correlation scores. We expect that we can obtain better performance by using a regression model trained on ASR transcripts, which can better model the impact of noisy data on the features. In our future work, we will build sample responses for each score category, tune the parameter values, and train the regression model all on ASR hypotheses. We hope this can solve the mismatch problem during training and evaluation, and can provide us even better correlation results.

6 Conclusion and Future Work

Most previous work on automated scoring of spontaneous speech used features mainly related to low-level information, such as fluency, pronunciation, prosody, as well as a few features measuring aspects such as vocabulary diversity and grammatical accuracy. In this paper, we focused on extracting content features to measure the speech proficiency in relatively higher-level aspect of spontaneous speech. Three features were computed to measure the similarity between a test response and a set of sample responses representing different levels of speaking proficiency. The similarity was calculated using different methods, including the lexical matching method VSM, and two corpus-based semantic similarity measures, LSA and PMI. Our experimental results showed that all the features obtained good correlations with human proficiency scores if there are no recognition errors in the text transcripts, with the PMI-based method performing the best over three similarity measures. However, if we used ASR transcripts, we observed a marked performance drop for the PMI-based method. Although we found that VSM and LSA were very robust to ASR errors, the overall correlations for the ASR condition were not as good as using human transcripts. To solve this problem, we proposed to use ASR confidence scores to improve the feature performance, and achieved similar results as when using human transcripts.

As we discussed in Section 5, all models were trained using human transcripts, which might decrease the performance when these models are applied directly to the ASR outputs. In our future work, we will compare models trained on human transcripts and on ASR outputs, and investigate whether we should use matching data for training and evaluation, or whether we should not introduce noise during training in order to maintain the validity of the models. We will also investigate whether the content features can provide additional information for automated speech scoring, and help build better scoring systems when they are combined with other non-content features, such as the features representing fluency, pronunciation, prosody, vocabulary diversity information. We will also explore generating other features measuring the higher-level aspects of the spoken responses. For example, we can extract features assessing the responses' relatedness to the stimulus of an opinion-based task. For contextual-based tasks, the test takers are asked to read or listen to some stimulus material, and answer a question based on this information. We can build models using these materials to check the correctness and relatedness of the spoken responses.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning, and Assessment*, 4(3):3–30.

Jared Bernstein, John De Jong, David Pisoni, and Brent Townshend. 2000. Two experiments on automatic scoring of spoken language proficiency. In *Proceedings of Integrating Speech Tech. in Learning (InSTIL)*.

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010a. Fluency and structural complexity as predictors of L2 oral proficiency. In *Proceedings of Interspeech*.

Jared Bernstein, Alistair Van Moere, and Jian Cheng. 2010b. Validating automated speaking tests. *Language Testing*, 27(3):355–377.

Lei Chen and Klaus Zechner. 2011a. Applying rhythm features to automatically assess non-native speech. In *Proceedings of Interspeech*.

Miao Chen and Klaus Zechner. 2011b. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of ACL-HLT*.

Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of NAACL-HLT*.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *IEEE Workshop on Auotmatic Speech Recognition and Understanding*.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107:989–999.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6):2862–2873.

Maxine Eskenazi, Abeer Alwan, and Helmer Strik. 2009. Spoken language technology for education. *Speech Communication*, 51(10):831–1038.

Peter W. Foltz, Darrell Laham, and Thomas K. Landauer. 1999. The Intelligent Essay Assessor: Applications to educational technology. *Interactive multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Horacio Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen. 2000. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(1-2):121–130.

Thomas K Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence*, September.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of ECML*.

Silke M. Witt and Steve J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(1-2):95–108.

Klaus Zechner and Xiaoming Xi. 2008. Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Klaus Zechner, Derrick Higgins, and Xiaoming Xi. 2007. Speechrater$^{\text{TM}}$: A construct-driven approach to score spontaneous non-native speech. In *Proceedings of the 2007 Workshop of the International Speech Communication Association (ISCA) Special Interest Group on Speech and Language Technology in Education (SLaTE)*.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.