# Application of Structural Events Detected on ASR Outputs for Automated Speaking Assessment

**Conference Paper** · September 2012

DOI: 10.21437/Interspeech.2012-236



**2 authors:**

Lei Chen
Rakuten Institute of Technology (RIT)
**75** PUBLICATIONS **1,403** CITATIONS
SEE PROFILE

Su-Youn Yoon
EduLab-inc
**47** PUBLICATIONS **692** CITATIONS
SEE PROFILE

# Application of Structural Events Detected on ASR Outputs for Automated Speaking Assessment

*Lei Chen, Su-Youn Yoon*

R&D, Educational Testing Service (ETS), Princeton, NJ, USA

LChen@ets.org, SYoon@ets.org

## Abstract

We investigated the features reflecting utterance structure and disfluency profile to improve the automated scoring of spontaneous speech responses by non-native speakers of English. On both human annotated structural events (SEs), e.g., clause structure and disfluencies, and automatically detected SEs on speech transcriptions, several features were derived and showed promisingly high correlations to the human proficiency scores. However, the usefulness of these SE-derived features on ASR hypotheses was still unknown. In this paper, we reported our studies related to the detection of SEs from noisy ASR outputs and the application of the detected SEs for automated speech scoring. We found that clause boundary (CB) detection was impacted much less compared to interruption point (IP) (of speech disfluencies) detection when facing ASR errors. Next, several features derived from the detected SEs were evaluated by considering their correlation to human scores and their relative importance in a linear regression model.

## 1. Introduction

Utterance structure and disfluencies (hesitation and dynamic error correction processes) are found to be important characteristics of spontaneous speech by previous psycholinguistic studies [1]. ESL (English as a second language) researchers actively investigate the relationships between these aspects and speech proficiency (e.g., [2]), and [3, 4] developed quantitative features. They have found that the utterance structure and disfluency based features can effectively and accurately assess speech proficiency.

Recently, a few studies have emerged in the automated speech assessment field which use the features related to utterance structure and disfluency profile, such as the complexity of sentences based on the average length of the clauses or sentences [5, 6], the parse-tree based features [7, 8], and the disfluency profile [9, 10]. However, most of such kind of research has been done on speech transcriptions and could not be directly used in operational speech scoring processes. As shown in [8], these measures could not show satisfactory empirical performance when extracted from the speech recognition outputs that generally contain errors. In addition, the disfluency features (e.g., [9, 10]) were limited to simple ones that do not consider utterance structure.

In this study, we present features reflecting both utterance structure and disfluency profile (focusing on pausing patterns) to improve the automated scoring of non-native spontaneous speech responses. The features were calculated in a fully automated way based on the automatically detected SEs from automatic speech recognition (ASR) outputs.

The paper is organized as follows: Section 2 reviews previous research; Section 3 reports on the data used in the paper; Section 4 describes the experiments on SE detection; Section 5 describes the application of the detected SEs for speech scoring purposes; Section 6 discusses the findings of our research and plans for future directions.

## 2. Previous Research

ESL researchers have developed many quantitative measures to estimate the utterance structure. Typical metrics for measuring syntactic complexity include: length of production units (e.g., T-units[1], clauses, verb phrases, and sentences), amount of embedding, subordination and coordination, range of structural types, and structural sophistication. [12, 6] found significant relationship between these features and students' proficiency levels.

Disfluency has been considered an important key to showing the sentence planning process. [13] classified disfluencies into two groups according to the locations within utterances: disfluencies that occurred at clause boundaries (hereafter, boundary disfluencies) and disfluencies that occurred within clauses (hereafter, within-clause disfluencies). [14, 15] found that boundary disfluencies serve as planning time, while within-clause disfluencies signal problems in sentence planning. In ESL research, [2] found a strong relationship between within-clause disfluencies and L2 speakers' proficiency. Speakers with low proficiency tend to pause within clauses more frequently than speakers with high proficiency. [4] showed that within-clause disfluency-based features have stronger correlations with human proficiency scores than the features only based on disfluencies.A combination of these structural elements, utterance structure and disfluency profile, can estimate the speakers' proficiency levels more accurately.

In speech processing, a large amount of research has been conducted to detect SEs in speech ASR outputs using both lexical and prosodic cues [16, 17, 18]. The detected SEs have been found to help many of the following natural language processing (NLP) tasks: speech parsing, information retrieval, machine translation, and extractive speech summarization [18]. Based on the improvements in automated SE detection, several studies in the area of automated speech scoring have emerged to investigate using the detected SEs reflecting utterance structure and disfluency profile for scoring spontaneous speech. [5] built statistical models to automatically detect clause boundaries (CBs) and interruption points (IPs) in speech transcriptions and found that the features computed from the detected SEs are promising for speech scoring. [7, 8] utilized parse-tree based syntactic complexity features on speech scoring tasks. However, most of

[1] A T-unit is defined as essentially a main clause plus any other clauses which are dependent upon it [11].

these studies were based on manual transcription and manual SE annotations. Therefore, the research findings could not be directly applied to the automated speech scoring that uses noisy ASR hypotheses. For example, [8] showed that the significant correlation between syntactic measures and speech proficiency (correlation coefficient = $0.49$) became insignificant when they were applied to the ASR hypotheses.

Our study provides new features based on both disfluency profile and utterance structures. This study is different from previous studies as follows: First, in contrast to the typical disfluency features, e.g., disfluency frequency and mean duration of disfluencies, the proposed features considered utterance structure and the validity of combining these two aspects was strongly supported by psycholinguistic and ESL research; Second, by actively employing speech technology (using ASR outputs and automated SE detection results), the proposed features could be fully generated automatically. Third, the proposed features were evaluated not only by the ordinary correlation analysis but also by comparing their relative importance with other widely used effective speech features.

# 3. Non-native Structural Event Corpus

Non-native speech data were collected from the TOEFL Practice [19]. In each TOEFL Practice test, test-takers were required to either provide information or express their opinions, based on personal experience or background knowledge. For example, the test-takers were asked for their opinions about living on or off campus.

A total of $1066$ responses was collected from examinees. Then, a group of experienced human raters scored these items $4$-point holistic scores based on the scoring rubrics designed for scoring the TOEFL Practice test. The speaking content was transcribed by a professional transcribing agency. On the transcriptions, SE annotations were added, including (1) locations of clause boundaries, (2) types of clauses (e.g., noun clauses, adjective clauses, adverb clauses, etc.), and (3) disfluencies.

For the research reported in this paper, we focus on two SEs: the locations of clause-ending boundaries (CBs) and interruption points (IPs) of disfluencies. Note that if several clauses (in different layers of a clause hierarchy) end at the same word boundary, these clause boundaries were collapsed into one CB event. More details about this data set, such as inter-rater annotation agreements, can be found in [20].

# 4. Structural Event Detection Experiments

## 4.1. Setup

In our experiment, the whole corpus described in Section 3 was split into a training set (*train*), a development test set (*dev*), and testing set (*test*), with no speaker overlap between any pair of sets. Table 1 summarizes the numbers of items and words, as well as the number of structural events in each dataset.

| | *train* | *dev* | *test* |
|---|---|---|---|
| # item | 664 | 101 | 301 |
| # word | 71036 | 10440 | 33516 |
| # CB | 6090 | 915 | 2837 |
| # IP | 1711 | 257 | 787 |

Table 1: The number of items, words, and structural events of the three sets in the TOEFL Practice corpus

In order to test the SE detection performance on ASR hypotheses, we also run speech recognition on both *dev* and *test* sets. A state-of-the-art Hidden Markov Model (HMM) speech recognizer was used in our experiments. In this ASR system, its acoustic model (AM) and language model (LM) were trained from about $750$ hours of non-native speech data and corresponding transcriptions. On the *test* set, our ASR resulted in a word error rate (WER) of $17.24\%$ for non-native learners' speech data, which was much smaller than the one reported in [8].

We followed [5] in using a Maximum Entropy (MaxEnt) model and a Conditional Random Field (CRF) model to classify each inter-word boundary in order to obtain its CB and IP type. The details of the experiments, including ordinary n-gram lexical features, the special lexical features for IP detection, and the implementations of these two models, can be found in [5].

For CB detection, we also used pause information after word-ending boundaries similar to [16]. A long pause after a word-ending boundary is an important feature indicating the existence of a clause/sentence boundary. To obtain such features, for transcription input, we used the P2FA forced alignment toolkit [21] to find inter-word pauses based on speech transcriptions. When examining ASR outputs, we can directly use the timing information output by the ASR system. In order to use the continuous-valued pause durations to be features for both the MaxEnt and CRF models, which require discrete features, we used the supervised discretization method in the Weka machine learning toolkit [22]. In particular, according to pause durations and corresponding CB labels, we used the method by Fayyad and Irani [23] to convert a continuous-valued pause duration to 6 distinct classes. We found that by adding such pause-related features, the CB detection accuracy of using both MaxEnt and CRF models has improved.

## 4.2. SE detection results

Since we treated the SE detection task as a binary classification task in this paper, we used standard evaluation metrics, including precision, recall, and $F1$ measurement. In order to evaluate SEs on ASR hypotheses, we used the NIST's SCTK2.4 package [24] to obtain the needed values, such as deletions and insertions of SEs, to compute these metrics. Table 2 summarizes the CB and IP detection results on two word input conditions: speech transcriptions vs. ASR hypotheses.

| Model | $F1_{Tran.}$ | $F1_{ASR}$ | $F1 \Downarrow (\%)$ |
|---|---|---|---|
| CB detection | | | |
| MaxEnt | 0.617 | 0.579 | 6.16 |
| CRF | 0.752 | 0.690 | 8.24 |
| IP detection | | | |
| MaxEnt | 0.412 | 0.301 | 26.9 |
| CRF | 0.410 | 0.304 | 25.9 |

Table 2: $F1$s for CB and IP detection tasks when using transcriptions vs. ASR outputs on the TOEFL Practice data

Compared to [5], since ASR outputs do not contain word fragments, we have removed all word fragments from our training transcriptions. As a result, losing such valuable information dramatically lowered IP detection performance on both transcriptions and ASR hypotheses input conditions. In addition, Table 2 clearly shows that errors in ASR hypotheses impact the IP detection task more than on the CB detection task.

## 5. Evaluation of the SE-derived features

Based on the SEs detected from speech responses, we then extracted several features for speech scoring purpose. Two features that were reported to have high correlations with human scores in [20, 5], i.e., IPC (IP frequency per clause) and IPW (IP frequency per word), were extracted. Furthermore, knowing clause structure allows us to utilize test-takers' pause patterns more effectively. Pauses can appear at clause boundaries or within clauses. We call the first type of pause a *boundary*-pause and the second type a *within-clause*-pause . Based on [4, 2], we anticipate that a *within-clause*-pause has a stronger link to speakers' proficiency levels than *boundary*-pause . Two features related to *within-clause*-pauses were implemented as follows:

- $aveDur$: average duration of all non-zero length *within-clause*-pauses .
- $ratioLP$: he ratio of long *within-clause*-pauses on all non-zero length *within-clause*-pauses . A long pause is determined when a pause's duration is longer than $0.190$ second, which is the median duration of *within-clause*-pauses .

### 5.1. Evaluation based on correlation analysis

On the *test* set, we produced CB and IP event sequences estimated by the MaxEnt and CRF models, respectively. Then, the SE-derived features were computed from five types of SE results on the *test* set, including (a) Human-Tran, human annotated SEs on manual transcriptions, (b) MaxEnt-Tran, the MaxEnt model's SE predictions on manual transcriptions, (c) CRF-Tran, the CRF model's SE predictions on manual transcriptions, (d) MaxEnt-ASR, the MaxEnt model's SE predictions on ASR hypotheses, (e) CRF-ASR, the CRF model's SE predictions on ASR hypotheses. Then, we computed Pearson correlation coefficients ($r$s) between these features and human holistic scores.

Table 3 reports the evaluation results of the features derived from the structural event estimations. First, we investigated the impact of only SE detection errors by comparing Human-Tran vs. MaxEnt-Tran and CRF-Tran. There were only small drops in $r_{IPC}$ and $r_{IPW}$. Furthermore, for $aveDur$ and $ratioLP$, the correlations increased. However, when using the machine-predicted SEs from noisy ASR outputs, we found large $r$ reduction from all of these SE-derived features except for $ratioLP$.

| SE resource | $r_{IPC}$ | $r_{IPW}$ | $r_{aveDur}$ | $r_{ratioLP}$ |
|---|---|---|---|---|
| Human-Tran | $-0.37$ | $-0.40$ | $-0.35$ | $-0.42$ |
| MaxEnt-Tran | $-0.37$ | $-0.38$ | $-0.43$ | $-0.46$ |
| CRF-Tran | $-0.31$ | $-0.30$ | $-0.36$ | $-0.49$ |
| MaxEnt-ASR | $-0.19$ | $-0.22$ | $-0.11$ | $-0.47$ |
| CRF-ASR | $-0.20$ | $-0.21$ | $-0.13$ | $-0.45$ |

Table 3: Correlation coefficients ($r$s) between the SE-derived features and human holistic scores

Both $r_{IPC}$ and $r_{IPW}$ were based on the IPs. The accuracy of IP detection decreased largely due to ASR errors and the increased errors finally resulted in significant correlation drops between IP-based proficiency features and human proficiency scores. On the contrary, the $r_{ratioLP}$ feature, based on the CBs that were relatively robust from ASR errors, showed slight or no drops in correlations.

### 5.2. Evaluation based on features' relative importances

In current speech scoring systems, a set of features has been widely used to cover many aspects of human speaking capabilities [25, 26]. Will the $ratioLP$ feature still be useful when used with those features? To answer this question, we extracted a set of widely used speech features following the methods described in [26]. Then, we built a linear regression model by using these features and the $ratioLP$ feature to predict human holistic scores. The relative importance of each feature in the regression model provides a way to answer the question. In our experiment, the speech features listed in Table 4 were compared with our SE-derived features.

| feature | description |
|---|---|
| *pace-word* | the number of words divided by the duration of the response |
| *pause* | the number of silences (the shortest silence needs to be at least $0.15$ seconds long) divided by the number of word |
| *prosody* | mean distance between two stressed syllables |
| *pronunciation* | the summation of alignment likelihoods divided by the number of letters of the recognized words, which is quite similar to the method widely used for pronunciation verification, such as the Goodness of Pronunciation (GOP) [25] |
| *language-use* | the language model score, which compares the language of non-native speech to a reference language model |

Table 4: A list of speech features that have been widely used on assessing spontaneous speech

We used the relaimpo R package [27] , a package designed for computing a series of relative importance measures for correlated features. In particular, we used the LMG measure [28], a metric named for its inventors (Lindeman, Merenda, and Gold), which is recommended by [27]. One of advantages of this metric is that all of the importance measures from all of the variables being investigated are $100\%$ and this makes it easy to analyze the variables's contributions. From Table 5, we can find that the *pace-word*, a speaking rate feature, plays an dominant role. The *pause*, *language-use*, and *pause* show relative importance of more than $10\%$. Our new *ratioLP* feature has a relative importance of $9.3\%$ when being used to predict human proficiency scores. Its importance is larger than a widely used feature, *pronunciation*, which has an importance of $6.3\%$. This suggests that the new *ratioLP* feature has a compatible performance with some currently used speech features and it can also enrich the feature's coverage on utterance structure and disfluency profile.

| Feature | LMG(%) |
|---|---|
| *pace-word* | 40.7 |
| *pause* | 17.9 |
| *language-use* | 13.6 |
| *prosody* | 12.2 |
| **ratioLP** | **9.3** |
| *pronunciation* | 6.3 |

Table 5: Features' LMG relative importance measures for predicting human proficiency scores

## 6. Discussion

Recently, several studies using entities about utterance structure and disfluency profile, which have been actively used in the ESL field, emerged in the automated speech scoring field. These studies were mainly conducted on error-free speech transcriptions. Although some positive results were reported, the actual usefulness on non-native spontaneous speech ASR outputs was still unknown. In this paper, we conducted several experiments on the usefulness of such SE-derived features for the automated speech scoring domain.

First, for SE detection tasks, we did not only use speech transcriptions like [5], but also used ASR hypotheses from a state-of-the-art non-native spontaneous speech recognizer, which has a much higher recognition accuracy than the recognizer used in [8]. We found that ASR errors heavily impacted IP detection performance. One reason for this is that word fragments appearing in speech transcriptions that provided very useful cues in many previous studies of disfluency detection were hard to obtain from ASR hypotheses. As for CB detection, the $F1$ only dropped about $6\%$ to $8\%$ when switching from speech transcriptions to ASR outputs.

Next, we extracted four SE-derived features, IPC and IPW, as suggested in [5] and two new features on *within-clause*-pauses . When using SEs detected from ASR outputs, we found that only our new $ratioLP$ feature still showed a promising correlation to human holistic scores. Furthermore, we also investigated the relative importance of the *ratioLP* feature when being used together with several widely used features for speech scoring. This analysis suggested that this new SE-derived feature is comparable to others.

For future research works, one direction will be to improve disfluency detection accuracy, especially for non-native spontaneous speech. For example, some prosodic cues will be considered. In addition, we will continue to explore other types of SEs, such as discourse markers, which play important roles in forming a coherent speech.

## 7. References


[1] M. Levelt, "Monitoring and self-repair in speech," *Cognition*, pp. 41–104, 1983.

[2] L. Temple, "Second language learner speech production," *Studia Linguistica*, pp. 288–297, 2000.

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

[4] G. J. Mizera, "Working memory and l2 oral fluency," Ph.D. dissertation, University of Pittsburgh, 2006.

[5] L. Chen and S. Yoon, "Detecting structural event for assessing non-native speech," in *6th Workshop on Innovative Use of NLP for Building Educational Applications*, 2011, p. 74.

[6] J. Bernstein, J. Cheng, and M. Suzuki, "Fluency and structural complexity as predictors of l2 oral proficiency," in *Interspeech 2010*, 2010.

[7] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[8] M. Chen and K. Zechner, "Computing and evaluating syntactic complexity features for automated scoring of spontaneous Non-Native speech," in *Proc. of ACL 2011*, 2011.

[9] 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.

[10] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, pp. 883–895, October 2009.

[11] K. W. Hunt, "Syntactic maturity in school children and adults," in *Monographs of the Society for Research in Child Development*. Chicago, IL: University of Chicago Press, 1970.

[12] N. Iwashita, "Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language," *Language Assessment Quarterly: An International Journal*, vol. 3, no. 2, pp. 151–169, 2006.

[13] F. Lounsbury, "Transitional probability, linguistic structure, and systems of habit-family hierarchies," *Psycholinguistics. A Survey of Theory and Research Problems*, pp. 93–101, 1954.

[14] D. S. Boomer, "Hesitation and grammatical encoding," *Language and Speech*, pp. 148–158, 1965.

[15] K. Bock and M. Levelt, "Language production: Grammatical encoding," in *Handbook of psycholinguistics*, M. Gernsbacher, Ed. San Diego: Academic Press, 1994, pp. 945–984.

[16] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcript," in *Proceedings of the International Speech Communication Association (ISCA) Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000*, 2000.

[17] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue University, 2004.

[18] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Woofers, "Speech segmentation and spoken document processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 59–69, May 2008.

[19] ETS, "TOEFL Practice Online Test (TPO)," 2006.

[20] L. Chen, J. Tetreault, and X. Xi, "Towards using structural events to assess non-native speech," in *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 2010, p. 74.

[21] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," in *Proc. of Acoustics*, 2008.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[23] U. Fayyad and K. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, pp. 87–102, 1992.

[24] NIST, *Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0 (Includes the SCLITE*. [Online]. Available: http://www.itl.nist.gov/iad/mig/tools/

[25] S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[26] L. Chen, K. Zechner, and X. Xi, "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech," in *NAACL-HLT*, 2009.

[27] U. Gramping, "Relative importance for linear regression in r: the package relaimpo," *Journal of Statistical Software*, vol. 17, no. 1, pp. 1–27, 2006.

[28] R. Lindeman, P. Merenda, and R. Gold, *Introduction to bivariate and multivariate analysis*. Scott, Foresman Glenview, IL, 1980.