Towards Using Structural Events To Assess Non-native Speech

Lei Chen, Joel Tetreault, Xiaoming Xi
Educational Testing Service (ETS)
Princeton, NJ 08540, USA
{LChen,JTetreault,XXi}@ets.org

Abstract

We investigated using structural events, e.g., clause and disfluency structure, from transcriptions of spontaneous non-native speech, to compute features for measuring speaking proficiency. Using a set of transcribed audio files collected from the TOEFL Practice Test Online (TPO), we conducted a sophisticated annotation of structural events, including clause boundaries and types, as well as disfluencies. Based on words and the annotated structural events, we extracted features related to syntactic complexity, e.g., the mean length of clause (MLC) and dependent clause frequency (DEPC), and a feature related to disfluencies, the interruption point frequency per clause (IPC). Among these features, the IPC shows the highest correlation with holistic scores ($r = -0.344$). Furthermore, we increased the correlation with human scores by normalizing IPC by (1) MLC ($r = -0.386$), (2) DEPC ($r = -0.429$), and (3) both ($r = -0.462$). In this research, the features derived from structural events of speech transcriptions are found to predict holistic scores measuring speaking proficiency. This suggests that structural events estimated on speech word strings provide a potential way for assessing non-native speech.

1 Introduction

In the last decade, a breakthrough in speech processing is the emergence of a lot of active research work on automatic estimation of structural events, e.g., sentence structure and disfluencies, on spontaneous speech (Shriberg et al., 2000; Liu, 2004; Ostendorf et al., 2008). The detected structural events have been successfully used in many natural language processing (NLP) applications (Ostendorf et al., 2008).

However, the structural events in speech data haven't been largely utilized by the research on using automatic speech recognition (ASR) technology to assess speech proficiency (Neumeyer et al., 2000; Zechner et al., 2007), which mainly used cues derived at the word level, such as timing information of spoken words. The information beyond the word level, e.g., clause/sentence structure of utterances and disfluency structure, has not been or is poorly represented. For example, in Zechner et al. (2007), only special words for filled pauses such as *um* and *uh* were obtained from ASR results to represent disfluencies.

Given the successful usage of structural events on a wide range of NLP applications and the fact that the usage of these events is missing in the automatic speech assessment research, a research question emerges: Can we use structural events of spontaneous speech to assess non-native speech proficiency?

We will address this question in this paper. The paper is organized as follows: Section 2 reviews previous research. Section 3 describes our annotation convention. Section 4 reports on the data collection, annotation, and quality control. Section 5 reports on features based on structural event annotations. Section 6 reports on our experiments. Section 7 discusses our findings and plans for future research work.

2 Previous Work

In the last decade, a large amount of research (Ostendorf et al., 2008) has been conducted on detection of structural events, e.g., sentence structure and disfluency structure, in spontaneous speech. In these research works, the structural events were detected with a quite high accuracy. Furthermore, the detected sentence and disfluency structures have been found to help many of the following NLP tasks, e.g., speech parsing, information retrieval, machine translation, and extractive speech summary (Ostendorf et al., 2008).

In the second language acquisition (SLA) and child language development research fields, the language development is measured according to fluency, accuracy, and complexity (Iwashita, 2006). The syntactic complexity of learners' writing data has been extensively studied in the SLA community (Ortega, 2003). Recently, this study has been extended to the learner's speaking data (Iwashita, 2006). Typical metrics for examining syntactic complexity include: length of production unit (e.g., T-unit, which is defined as essentially a main clause plus any other clauses which are dependent upon it (Hunt, 1970), clauses, verb phrases, and sentences), amount of embedding, subordination and coordination, range of structural types, and structure sophistication.

Iwashita (2006) investigated several measures for syntactic complexity on the data from learners of Japanese. The author reported that some measurements, e.g., T-unit length, the number of clauses per T-unit, and the number of independent clauses per T-Unit, were good at predicting learners' proficiency levels.

In addition, some previous studies used measurements related to disfluencies to assess speaking proficiency. For example, Lennon (1990) used a dozen features related to speed, pauses, and several disfluency markers, such as filler pauses per T-unit, to measure four German-speaking women's English improvement during a half year study in England. He found a significant change in filled pauses per T-unit during the studying process.

The features related to syntactic complexity and the features related to "smoothness" (disfluency) of speech were jointly used in some previous studies. For example, Mizera (2006) used fluency factors related to speed, voiced smoothness (frequencies of repetitions or self-corrections), pauses, syntactic complexity (mean length of T-units), and accuracy, to measure speaking proficiency on 20 non-native English speakers. In this experiment, disfluency-related factors, such as total voiced disfluencies, had a high correlation with fluency ($r = -0.45$). However, the syntactic complexity factor only showed a moderate correlation ($r = 0.310$). Yoon (2009) implemented an automated disfluency detection method and found that the disfluency-related features lead to the moderate improvement in the automated speech proficiency scoring.

There were limitations on using the features reported in these SLA studies on standard language tests. For example, only a very limited number of subjects (from 20 to 30 speakers) were used in these studies. Second, the speaking content was narrations of picture books or cartoon videos rather than standard test questions. Therefore, we conducted a study using a much larger data set obtained from real speech tests to address these limitations.

3 Structural Event Annotation Convention

To annotate structural events of speech content, we have developed a convention based on previous studies and our observations on the TOEFL Practice Online (TPO) test data. Defining clauses is a relatively simple task; however, defining clause boundaries and specifying which elements fall within a particular clause is a much more challenging task for spoken discourse, due to the presence of grammatical errors, fragments, repetitions, self corrections, and conversation fillers.

Foster et al. (Foster et al., 2000) review various units for analyzing spoken language, including syntactic, semantic and intonational units, and propose a new analysis of speech unit (AS-Unit) that they claim is appropriate for many different purposes. In this study, we focused on clauses given the characteristics of spontaneous speech. Also, we defined clause types based on grammar books such as (Azar, 2003). The following clause types were defined:

- Simple sentence (**SS**) contains a subject and a verb, and expresses a complete thought.

- Independent clause (**I**) is the main clause that can stand along syntactically as a complete sentence. It consists minimally a subject and a finite verb (a verb that shows tense, person, or singular/plural, e.g., *he goes*, *I went*, and *I was*).

- Subordinate clause is a clause in a complex sentence that cannot stand alone as a complete sentence and that functions within the sentence as a noun, a verb complement, an adjective or an adverb. There are three types of subordinate clauses: noun clause (**NC**), relative clause that functions as an adjective (**ADJ**), adverbial clause that functions as an adverb (**ADV**).

- Coordinate clause (**CC**) is a clause in a compound sentence that is grammatically equivalent to the main clause and that performs the same grammatical function.

- Adverbial phrase (**ADVP**) is a separate clause from the main clause that contains a non-finite verb (a verb that does not show tense, person, or singular/plural).

The clause boundaries and clause types were annotated on the word transcriptions. Round brackets were used to indicate the beginning and end of a clause. Then, the abbreviations described above for clause types were added. Also, if a specific boundary serves as the boundaries for both the local and global clause, the abbreviation of the local clause was followed by that of the global. Some examples of clause boundaries and types are reported in Table 1.

In our annotation manual, a speech disfluency contains three parts:

- *Reparandum*: the speech portion that will be repeated, corrected, or even abandoned. The end of the reparandum is called the interruption point (**IP**), which indicates the stop of a normal fluent speech stream.

- *Editing phrase*: optional inserted words, e.g., *um*.

- *Correction*: the speech portion that repeats, corrects, or even starts new content.

In our annotation manual, the reparandum was enclosed by "*", the editing phrase was enclosed by "%", and the correction was enclosed by "$". For example, in the following utterance, "He is a * very mad * % er % $ very bad $ cop", "very mad" was corrected by "very bad" and an editing phrase, *er*, was inserted.

4 Data Collection and Annotation

4.1 Audio data collection and scoring

About 1300 speech responses from the TPO test were collected and transcribed. Each item was scored by two experienced human raters independently using a 4-point holistic score based on the scoring rubrics designed for the test.

In the TPO test, some tasks required test-takers to provide information or opinions on familiar topics based on their personal experience or background knowledge. Others required them to summarize and synthesize information presented in listening and/or reading materials. Each test-taker was required to finish six items in one test session. Each item has a 45 or 60 seconds response time.

4.2 Annotation procedure

Two annotators (who were not the human raters mentioned above) with a linguistics background and past linguistics annotation experience were first presented with a draft of the annotation convention. After reading through it, the annotators, as well as the second and third author completed four iterative loops of rating 4 or 5 responses per meeting. All four discussed differences in annotations and the convention was refined as needed. After the final iteration of comparisons, the raters seemed to have very few disagreement and thus began annotating sets of responses. Each set consisted of roughly 50-75 responses and then a kappa set of 30-50 responses which both annotators completed. Accordingly, between the two annotators, a set comprised roughly 130 to 200 responses. Each response takes roughly 3-8 minutes to annotate. The annotators were instructed to listen to the corresponding audio file if they needed the prosodic information to annotate a particular speech disfluency event.

Clause type	Example
SS	(That's right |SS)
I	(He turned away |I) as soon as he saw me |ADV)
NC	((What he did |NC) shocked me |I)
ADJ	(She is the woman (I told you about |ADJ)|I)
ADV	(As soon as he saw me |ADV) (he turned away |I)
CC	(I will go home |I) (and he will go to work |CC)
ADVP	(While walking to class |ADVP) (I ran into a friend |I)

Table 1: Examples of clause boundary and type annotation

4.3 Evaluation of annotation

To evaluate the quality of structural event annotation, we measured the inter-rater agreement on clause boundary (CB) annotation and interruption point (IP) of disfluencies[1].

We used Cohen's κ to calculate the annotator agreement on each kappa set. κ is calculated on the absence or presence of a boundary marker (either a clause boundary (CB) or an interruption point (IP) between consecutive words). For each consecutive pair of words, we check for the existence of one or more boundaries, and collapse the set into one term "boundary" and then compute the agreement on this reduced annotation.

In Table 2, we list the annotator agreement for both boundary events over 4 kappa sets. The second column refers to the number of speech responses in the kappa set, the next two columns refer to the annotator agreement using the Cohen's κ value on CB and IP annotation results.

Set	N	κ CB	κ IP
Set1	54	0.886	0.626
Set2	71	0.847	0.687
Set3	35	0.855	0.695
Set4	34	0.899	0.833

Table 2: Between-rater agreement of structural event annotation

In general, a κ of 0.8-1.0 represents excellent agreement, 0.6-0.8 represents good agreement, and so forth. Over each kappa set, κ for CB annotations ranges between 0.8 and 0.9, which is an excellent agreement; κ for IP annotation ranges between 0.6 and 0.8, which is a good agreement. Compared to annotating clauses, marking disfluencies is more challenging. As a result, a lower between-rater agreement is expected.

[1]Measurement on CBs and IPs can provide a rough quality measurement of annotations. In addition, doing so is more important to us since automatic detection of these two types of events will be investigated in future.

5 Features Derived On Structural Events

Based on the structural event annotations, including clause boundaries and their types, as well as disfluencies, some features measuring syntactic complexity and disfluency profile were derived.

Since simple sentence (SS), independent clause (I), and conjunct clause (CC) represent a complete idea, we treat them as an approximate to a T-unit (T). The clauses that have no complete idea, are dependent clauses (DEP), including noun clauses (N), relative clauses that function as adjective (ADJ), adverbial clauses (ADV), and adverbial phrases (ADVP). The total number of clauses is a summation of the number of T-units (T), dependent clauses (DEP), and fragments[2] (denoted as F). Therefore,

$$\begin{aligned} N_T &= N_{SS} + N_I + N_{CC} \\ N_{DEP} &= N_{NC} + N_{ADJ} + N_{ADV} + N_{ADVP} \\ N_C &= N_T + N_{DEP} + N_F \end{aligned}$$

Assuming N_w is the total number of words in the speech response (without pruning speech repairs), the following features, including mean length of clause (MLC), dependent clauses per clause (DEPC), and interruption points per clause (IPC), are derived:

$$MLC = N_w/N_C$$

[2]It is either a subordinate clause that does not have a corresponding independent clause or a string of words without a subject or a verb that does not express a complete thought.

$$DEPC = N_{DEP}/N_C$$
$$IPC = N_{IP}/N_C$$

Furthermore, we elaborated the IPC feature. Disfluency is a complex behavior and is influenced by a variety of factors, such as proficiency level, speaking rate, and familiarity with speaking content. The complexity of utterances is also an important factor on the disfluency pattern. For example, Roll et al. (Roll et al., 2007) found that complexity of expression computed based on the language's parsing tree structure influenced the frequency of disfluencies in their experiment on Swedish. Therefore, since disfluency frequency was not only influenced by the test-takers' speaking proficiency but also by the utterance's syntactic structure's difficulty, we reduced the impact from the syntactic structure so that we can focus on speakers' ability. For this purpose, we normalized IPC by dividing by some features related to syntactic-structure's complexity, including MLC, DEPC, and both. Therefore, the following elaborated disfluency-related features were derived:

$$IPCn1 = IPC/MLC$$
$$IPCn2 = IPC/DEPC$$
$$IPCn3 = IPC/MLC/DEPC$$

6 Experiment

For each item, two human raters rated it separately with a score from 1 to 4. If these two scores are consistent (the difference between two scores is either zero or one), we put this item in an item-pool. Finally, a total of $1,257$ audio items were included in the pool. Following the score-handling protocol used in the TPO test, we used the first human rater's score as the item score. From the obtained item-pool, we selected speakers with more than three items so that the averaged score per speaker can be estimated on several items to achieve a robust score computation[3]. As a result, 175 speakers[4] were selected.

[3]The mean holistic score of these speakers is 2.786, which is close to the mean holistic score of the selected item-pool (2.785), indicating that score distribution was kept after focusing on speakers with more than three items.

[4]If a speaker was assigned in a Kappa set in the annotation as described in Section 4, this speaker would have as many as 12 annotated items. Therefore, the minimum number of speakers from the item-pool was about 105 ($1257/12$).

For each speaker, his or her annotations of words and structural events were used to extract the features described in Section 5. Then, we computed the Pearson correlation among the obtained features with the averaged holistic scores per speaker.

Feature	r
MLC	0.211
DEPC	0.284
IPC	-0.344
IPCn1	-0.386
IPCn2	-0.429
IPCn3	-0.462

Table 3: Correlation coefficients (rs) between the features derived from structural events with human scores averaged on test takers

Table 3 reports on the correlation coefficient (r) between the proposed features derived from structural events with holistic scores. Relying on three simple structural event annotations, i.e., clause boundaries, dependent clauses, and interruption points in speech disfluencies, some promising correlations between features with holistic scores were found. Between the two syntactic complexity features, the DEPC has a higher correlation with holistic scores than the MLC ($0.284 > 0.211$). It appears that a measurement about clauses' embedding profile is more informative about a speaker's proficiency level. Second, compared to features measuring syntactic complexity, the feature measuring the disfluency profile is better to predict human holistic scores on this non-native data set. For example, IPC has a r of -0.344, which is better than the features about clause lengths or embedding. Finally, by jointly using the structural events related to clauses and disfluencies, we can further achieve a further improved r. Compared to IPC, IPCn3 has a relative 34.30% correlation increase. This is consistent with our idea of reducing utterance-complexity's impact on disfluency-related features.

7 Discussion

In most current automatic speech assessment systems, features derived from recognized words, such as delivery features about speaking rate, pause information, and accuracy related to word identities, have been widely used to assess non-native speech from

fluency and accuracy points of view. However, information beyond recognized words, e.g., the structure of clauses and disfluencies, has only received limited attention. Although several previous SLA studies used features derived from structural events to measure speaking proficiency, these studies were limited and the findings from them were difficult to directly apply to on large-scale standard tests.

In this paper, using a large-sized data set collected in the TPO speaking test, we conducted an sophisticated annotation of structural events, including boundaries and types of clauses and disfluencies, from transcriptions of spontaneous speech test responses. A series of features were derived from these structural event annotations and were evaluated according to their correlations with holistic scores. We found that disfluency-related features have higher correlations to human holistic scores than features about syntactic complexity, which confirms the result reported in (Mizera, 2006). In spontaneous speech utterances, simple syntactic structure tends to be utilized by speakers. This is in contrast to sophisticated syntactic structure appearing in writing. This may cause that complexity-related features are poor at predicting fluency scores. On the other hand, disfluencies, a pattern unique to spontaneous speech, were found to play a more important role in indicating speaking proficiency levels.

Although syntactic complexity features were not highly indicative of holistic scores, they were useful to further improve disfluency-related features' correlation with holistic scores. By normalizing IPC using measurements representing syntactic complexity, we can highlight contributions from speakers' proficiency levels. Therefore, in our experiment, IPCn3 shows a 34.30% relative improvement in its correlation coefficient with human holistic scores over the original IPC.

The study reported in this paper suggests promise that structural events beyond speech recognition results can be utilized to measure non-native speaker proficiency levels. Recently, in the NLP research field, an increasing amount of effort has been made on structural event detection in spontaneous speech (Ostendorf et al., 2008). Therefore, such progress can benefit the study of automatic estimation of structural events on non-native speech data.

For our future research plan, first, we will investigate automatically detecting these structural events from speech transcriptions and recognition hypotheses. Second, the features derived from the obtained structural events will be used to augment the features in automatic speech assessment research to provide a wider construct coverage than fluency and pronunciation features do.

References

B. Azar. 2003. *Fundamentals of English grammar*. Pearson Longman, White Plains, NY, 3rd edition.

P. Foster, A. Tonkyn, and G. Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354.

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

N. Iwashita. 2006. Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly: An International Journal*, 3(2):151–169.

P. Lennon. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3):387–417.

Y. Liu. 2004. *Structural Event Detection for Rich Transcription of Speech*. Ph.D. thesis, Purdue University.

G. J. Mizera. 2006. *Working memory and L2 oral fluency*. Ph.D. thesis, University of Pittsburgh.

L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. 2000. Automatic Scoring of Pronunciation Quality. *Speech Communication*, 30:83–93.

L. Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4):492.

M. Ostendorf et al. 2008. Speech segmentation and spoken document processing. *Signal Processing Magazine, IEEE*, 25(3):59–69, May.

M. Roll, J. Frid, and M. Horne. 2007. Measuring syntactic complexity in spontaneous spoken Swedish. *Language and Speech*, 50(2):227.

E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154.

S. Yoon. 2009. *Automated assessment of speech fluency for L2 English learners*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

K. Zechner, D. Higgins, and Xiaoming Xi. 2007. SpeechRater: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech. In *Proc. SLaTE*.