A COMPARISON OF ASR AND HUMAN ERRORS FOR TRANSCRIPTION OF NON-NATIVE SPONTANEOUS SPEECH

Matthew Mulholland, Melissa Lopez, Keelan Evanini, Anastassia Loukina, Yao Qian

Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
{mmulholland, mlopez002, kevanini, aloukina, yqian}@ets.org

ABSTRACT

In this paper, we compare ASR and human transcriptions of non-native speech to investigate to what extent the accuracy and the patterns of errors of a modern ASR system match those of human listeners in the context of automated assessment of L2 English language proficiency. We obtained multiple naïve transcriptions of short fragments of non-native spontaneous speech with different proficiency levels using crowdsourcing and matched these against the output of an ASR system. We compare WER and recall at the fragment level and consider human-ASR agreement at the word level. We find that we are able to attain a commensurate level of transcription quality using ASR, but the patterns of errors between the two groups differ at the word level.

Index Terms— automatic speech recognition, speech transcription, L2 speech, crowdsourcing

1. INTRODUCTION

Automatic recognition of non-native speech is a challenging task due to the large number of potential deviations from native speaker norms at all linguistic levels, ranging from pronunciation to vocabulary choice and syntax. It is not surprising, therefore, that the accuracy of ASR for non-native speakers is typically substantially lower than for native speakers [1]. Yet understanding non-native speech can also be a difficult task for human listeners and studies in human transcription have also shown lower agreement between both expert and naïve transcribers [2, 3]. In this paper we compare ASR and human transcriptions of non-native spontaneous speech to investigate to what extent the accuracy and the patterns of errors of a modern ASR system match those of human listeners.

State-of-the-art GMM-based ASR systems typically achieve a WER in the range of 18-23% for large vocabulary continuous speech recognition for native speech, depending on the corpus [4, 5], while WER for DNN-based systems can be as low as 6.5% [5]. However, ASR accuracy for the task of large vocabulary spontaneous speech tends to be substantially lower for non-native speakers, with WER values around 30-35% for GMM-based systems [6, 1]. This is due to the fact that non-native speech can deviate from native speech in many ways, including patterns of pronunciation, grammar, syntax, and disfluencies. Therefore, a substantial amount of research has focused on improving ASR performance on non-native speech [7]. Recently, [8] reported that a DNN-based speaker dependent system achieved a WER of 20% on non-native spontaneous speech drawn from the domain of large-scale English proficiency assessment.

The relatively low accuracy of ASR for non-native speech is a problem for applications designed with non-native speakers in mind, such as automated scoring of spoken language proficiency, since the ASR output of the non-native speaker's response is used to compute various features for evaluating non-native spoken language proficiency [9, 10]. Low ASR accuracy has a detrimental effect on the performance of such scoring systems [11], and also raises questions about the validity of automated scores which are based on an inaccurate ASR hypothesis. Transcription of non-native speech is also challenging for human listeners—while transcribers can achieve WER values as low as 2-5% for transcription of spontaneous native speech [12, 13], the accuracy tends to be substantially lower for non-native spontaneous speech. For example, [2, 3] reported WER of 15-20% for expert transcribers, and the agreement is even lower for naïve listeners: [3, 14] reported a WER of about 30% for transcriptions obtained via Amazon Mechanical Turk. [1]

Thus, on the corpus level, the accuracy of a standard GMM-based ASR system on non-native spontaneous speech appears to be comparable to that of naïve human listeners (around 30% WER). The question that one may ask at this point is how different are the patterns of errors made by humans and the ASR system? Previous studies which compared human and ASR transcriptions of native data identified both similarities and differences in error patterns: for example, [16] proposed that doubly confusable pairs, or words that are both acoustically similar and have similar language model probabilities, may explain some recognition errors for both humans and ASR systems. [17] found that humans were better able to correctly identify words in isolation than an ASR system; however, an ASR system with a trigram language model outperformed humans who were given one or two words of context. To our knowledge, no such studies have been done for non-native speech. Yet, a better understanding of the patterns of errors made by human listeners and ASR systems when transcribing non-native speech could lead to a better understanding of the impact that ASR accuracy has on speech applications for non-native speakers, such as systems for computer-assisted language learning. In this paper, we therefore compare transcriptions from naïve listeners collected using Amazon Mechanical Turk and the output of a GMM-based ASR system to address the following research questions: (a) do human listeners and ASR achieve the same accuracy? and (b) are the words misrecognized by human

[1] A similar discrepancy exists between transcription accuracy rates for native vs. non-native restricted speech, although the absolute magnitude of the difference is smaller. For example, [15] reported on the results of transcribing native and non-native speech produced while providing route instructions to a robot; in this restricted vocabulary transcription task, the WER was 3.6% for native speech and 6.4% for non-native speech.

listeners the same as the words misrecognized by the ASR? (c) how do confidence scores computed by the ASR correlate with the number of transcribers who were able to recognize a given word?

2. DATA AND METHODOLOGY

The study is based on a corpus of non-native unscripted English speech which contains 143 responses to a large-scale assessment of English language proficiency for academic purposes collected from 140 non-native speakers of seven different native languages. The speakers were asked to respond to a prompt about the content of a conversation or a written text and given one minute to record their responses. For more information on the corpus, see [14].

We first obtained orthographic transcriptions for all 143 spoken responses. These transcriptions were done by a *professional transcription agency*. Our goal was to obtain as accurate a transcription of each response as possible. Therefore, for all responses transcribers had access to the full length of the audio file as well as background information about the item such as the source text that the test-takers were asked to summarize.

We then collected the *crowdsourced transcriptions* using Amazon Mechanical Turk (MTurk). Unlike expert transcribers, the transcribers recruited via MTurk (hereafter referred to as Turkers) only had access to limited information about each response to ensure a fairer comparison between ASR and human transcriptions. Each response was split into several fragments of approximately 8 words in length which were presented to the Turkers in randomized order. The final set consisted of 1,149 audio fragments and their accompanying transcriptions. We collected 5 transcriptions for each fragment for a total of 5,745 short transcriptions. See [14] for more information on the fragment selection and data collection procedures.

To split the responses into fragments, we used clause boundaries [18] along with punctuation from the orthographic transcriptions and pauses identified by The Penn Phonetics Lab Forced Aligner [19] to split the recordings and reference transcriptions into short fragments of 5-13 words in length.

We limited the Turkers to those with addresses in the United States who completed a short qualification test and demonstrated a good-faith effort. After initial data collection, we applied statistical analyses to identify and exclude the Turkers whose responses were significantly different from the others. We obtained new annotations as necessary so that the total number of Turkers for each fragment was 5. The results we present in this paper only include the Turkers whose responses were used for the analysis.

Finally, we obtained *ASR output* for each of the responses and then split each response into fragments that corresponded to the reference transcription fragments. To obtain the ASR output, a third-party ASR system was optimized for non-native speech using a proprietary training corpus consisting of over 800 hours of non-native spontaneous speech from the same assessment that the corpus of 143 responses was drawn from (with no speaker overlap). The ASR system used a GMM-based crossword triphone acoustic model and a 4-gram language model with a vocabulary size of approximately 65,000 words. In order to compare the ASR output to both the reference transcriptions and the crowdsourced transcriptions, we separated the output for each response into fragments based on the timestamps where the audio was split previously. In some cases, there were conflicts between the timestamps of the split fragments and the timestamps of the ASR output. In this study, we only included ASR output words that were completely within the boundaries of the original fragments.

We compared WER and recall for ASR and Turker transcriptions for each fragment to evaluate the overall accuracy of the two systems as well as the agreement at the word level. Finally, we computed correlations between word confidence scores returned by the recognizer and the number of Turkers who correctly transcribed each word. For the crowdsourced data, we used two approaches to combining the Turker transcriptions of each fragment. We also used the ROVER method [20] to merge the five Turker transcriptions for each fragment into a single string. We refer to this later as the "Turker merge" method.

The full corpus used for the analysis presented in this paper consisted of 1,149 fragments which included 9,494 words. We also explored the effect of stemming and the exclusion of function words. For stemming, we recreated each set of full transcriptions with stemmed versions of each word to compute transcription accuracy without penalizing relatively minor differences in morphology (Stem). In order to explore the relationship between recognition accuracy and word types, we created a subset of words for each transcription consisting only of content words (Content). After filtering non-content words (and removing empty fragments), we were left with a subset of 1,131 fragments and 3,641 content words. Finally, we applied both processing configurations (function word exclusion and stemming) to the data for a third iteration of our analyses (StemContent).

3. RESULTS

3.1. Fragment-level comparisons

3.1.1. WER-based accuracy

For our first analysis, we computed the word error rate (WER) using the set of reference transcriptions. For the crowdsourced data, we first calculated a mean WER for each fragment by averaging the WER of all five transcribers for that fragment ("Turker mean"). Separately, we calculated the WER for each fragment by comparing the merged Turker transcription to the reference transcription ("Turker merge"). For the ASR output, we calculated the WER for each fragment using the reference transcription.

We found that the average WER across Turkers across fragments was 26.4% using the "Turker mean" method and 18.5% using the "Turker merge" method. The average WER across fragments for the ASR system was 30.8%. The relationship between the ASR-Turker fragment-level word error rates using Pearson's correlation was r = 0.57 (Turker mean) and r = 0.54 (Turker merge). All correlations are significant at α=0.0001. We performed the same analysis using stemmed, content word, and stemmed content word data sets and found that stemming led to a decrease in WER for both methods, but WER increased when the data set was restricted to content words only. The agreement between ASR and Turkers was also lower for the data set restricted to content words. Full results are shown in Table 1.

3.1.2. Recall-based accuracy

We also calculated the recall for the ASR output and the crowdsourced transcriptions at the fragment level. To compute this, each word in the reference transcription that was aligned to the identical word in the ASR or crowdsourced transcription received a "recognized" score of 1; words in the reference transcription that were not aligned to a matching word, i.e., substitutions and deletions, received a "recognized" score of 0. Words that were inserted in the ASR output or the crowdsourced transcriptions were not considered for the

recall-based analysis. Instead of penalizing a transcription for insertions, recall simply demonstrates the percentage of words that were present in the reference transcription that were correctly aligned to words in the other transcription (either ASR or Turker).

The average recall across Turkers across fragments was 76.5% using the Turker mean method and 83.6% using the Turker merge method. The average recall across fragments for the ASR system was 72.6%. The correlation between fragment-level recall between as ASR and the two Turker methods was r = 0.51 (Turker mean) and r = 0.49 (Turker merge). Both correlations are significant at α = 0.0001. We performed the same analysis using stemmed and content word-only data sets. As in the case of WER, stemming led to an increase in recall. Unlike WER, removing function words did not seem to have any substantial effect on recall for the Turker mean. Full results are shown in Table 1.

Method	Data set	WER	r_{wer}	Recall	r_{rec}
ASR	Original	30.8	–	72.6	–
	Stem	29.3	–	74.0	–
	Content	31.1	–	74.1	–
	StemContent	27.5	–	77.6	–
Turker mean	Original	26.4	0.57	76.5	0.51
	Stem	24.5	0.57	78.3	0.51
	Content	31.8	0.54	76.9	0.41
	StemContent	28.8	0.55	79.8	0.40
Turker merge	Original	18.5	0.54	83.6	0.49
	Stem	17.7	0.54	84.3	0.48
	Content	23.4	0.50	83.3	0.30
	StemContent	21.7	0.49	84.9	0.37

Table 1. Fragment-level results. For each version of the data set (Data set), the table shows the average WER and recall across fragments for ASR and Turker aggregation methods (Turker mean and Turker merge). In the rows for Turkers, the table shows the correlation between ASR and the given Turker method for WER (r_{wer}) and recall (r_{rec}) at the fragment level. All correlations are significant at α=0.0001.

3.2. Word-level comparisons

We also compared the ASR output directly with the crowdsourced transcriptions using word-level statistics. As described in Section 3.1.2, we assigned a "recognized" value (1 or 0) to each word in the reference transcription for each set of data. For the word-level analysis, we only use the Turker mean method, as this provides more information about the number of Turkers who recognized each word. Since we obtained transcriptions from five Turkers, we were also able to create a human recognition score value between 0 and 5 for each word, where 5 means that all 5 Turkers recognized the word and 0 means that no Turkers recognized the word. We used these values to compare the ASR word accuracy to the groups of Turkers who transcribed the same word.

3.2.1. Recognition reliability

We calculated the agreement on word recognition between Turkers and ASR based on words in the reference transcriptions. For the ASR system, we used each word's "recognized" value. For Turkers, we used the majority "recognized" value for each word. With five Turkers transcribing each word, the majority value would be the "recognized" value shared by at least three Turkers. For example, if three out of five Turkers recognized a word then the majority value would be 1, or "recognized".

We first created confusion matrices comparing the number of words recognized by the ASR system and by the majority of (three or more) Turkers. These are shared in Table 2 for the full set (Original) and in Table 3 for the set containing stemmed content words (StemContent). The largest group in both matrices contains words that were recognized by both ASR and Turkers, followed by words recognized by Turkers but unrecognized by ASR. The ASR system recognized 81% of words recognized by Turkers in the Original set and 84% in the StemContent set. Additionally, close to 45% of words in the Original set and 50% in the StemContent set not recognized by Turkers were recognized by the ASR system.

	ASR rec.	ASR unrec.
Turker rec.	6,168	1,491
Turker unrec.	822	1,013

Table 2. Confusion matrix for the Original set containing 9,494 words. This table compares the number of words recognized and unrecognized by the majority of Turkers (Turker rec. and Turker unrec.) and by ASR (ASR rec. and ASR unrec.)

	ASR rec.	ASR unrec.
Turker rec.	2,554	496
Turker unrec.	294	297

Table 3. Confusion matrix for the StemContent set containing 3,641 words. This table compares the number of words recognized and unrecognized by the majority of Turkers (Turker rec. and Turker unrec.) and by ASR (ASR rec. and ASR unrec.)

We then calculated the raw agreement for each data set. The raw agreement between ASR and Turkers at the word-level was 75.7%. In agreement with fragment-level results, the agreement was higher when the transcriptions were stemmed. However, agreement increased after removing the function words, unlike the fragment-level results. Table 4 shows the agreement statistics for all sets.

Method	Data set	Raw agreement
Turker mean	Original	75.7
	Stem	77.3
	Content	77.1
	StemContent	78.4

Table 4. Word-level results. For each version of the data set (Data set), the table shows the raw agreement between the ASR and the Turker mean at the word level.

Finally, we investigated whether the total number of Turkers who recognized the word was related to the ASR accuracy for this word. We used the human recognition score to determine the average number of human transcribers who recognized a word correctly when the ASR system recognized or did not recognize the word. For cases where the ASR system correctly recognized a word, the average human recognition score was 4.19 (SD = 1.25). For cases where the ASR system did not recognize a word, the average human recognition score was 2.86 (SD = 1.91). This means that on average, 4 out of 5 Turkers recognized a word that ASR could recognize, while words unrecognized by ASR were usually recognized by about half (2-3) of the Turkers. The results are similar across data sets.

We also performed the reverse of this analysis – based on the number of Turkers who recognized a word, how often did the ASR system recognize that word? As expected, the ASR system recognized a smaller percentage of words as the human recognition score decreased. For example, the ASR system recognized 4,133 of the 4,687 words that all 5 Turkers recognized, but it only recognized 167 of the 665 words that no Turkers recognized. This pattern is similar when words are stemmed and function words are removed. See Table 5 for a full list of this information.

Human rec.	N_w Turkers	N_w ASR	% words
5	4,867	4,133	84.9
4	1,893	1,438	76.0
3	899	597	66.4
2	655	405	61.8
1	515	250	48.5
0	665	167	25.1

Table 5. Number of words (N_w Turkers) and percentage of words (% words) recognized by ASR (N_w ASR) based on human recognition score (Human rec.).

3.2.2. Confidence scores

The ASR system produces a confidence score for each word. We viewed the human recognition score as somewhat of an analog to this value in the crowdsourced data. We explored the relationship between these word-level confidence scores and found a fairly weak correlation of r = 0.34, $p < 0.0001$. The correlation was similar for stemmed sets but increased slightly to r = 0.37 ($p < 0.0001$) for the unstemmed content word set (Content). The relationships between the human recognition score and the ASR language model and acoustic model scores taken individually were much weaker– r = 0.14, $p < 0.00001$ and r = 0.02, $p < 0.02$, respectively.

We performed this analysis again by averaging word-level ASR scores and human recognition scores for each fragment to analyze their relationship at the fragment level. The relationship between the confidence score and human recognition score at the fragment level was stronger than at the word level, but it was still fairly weak: r = 0.46, $p < 0.00001$. The fragment-level relationship was higher than at the word level for all sets, but r was somewhat lower when function words were removed. The relationship using the language model and acoustic model scores show a similar pattern: r = 0.20 and r = 0.13, respectively. All correlations in this part of the analysis were significant with $p < 0.00001$.

3.3. Discussion

In this study, we compared the accuracy of groups of naïve humans with that of ASR output in transcribing spontaneous speech by non-native English speakers. Our aim was partially to demonstrate the validity of automated scoring, but also to provide further analysis of ASR error biases. Our results show that at the corpus level, the ASR system and an averaged group of Turkers performed within 5% of each other on both WER and recall measurements when compared to expert transcribers. The WER for ASR and Turkers was also consistent with what has been reported in previous studies.

We also found moderate agreement in accuracy between the two groups at the fragment level with the correlation between ASR and Turker r=0.57 for WER and r = 0.51 for recall. In other words, fragments that were more difficult for human transcribers were also more difficult for ASR. The two transcription methods differ more significantly at the word level.

We should note here that the pattern of agreement between ASR and Turkers closely matches the agreement between the Turkers. [14] reported high correlation in the accuracy between the Turkers at the fragment level with r = 0.82, but relatively low agreement at word level with Fleiss's κ = 0.429. However, we found that the correlation between ASR and Turkers was lower at both the fragment level and word level. We should additionally note that merging the Turkers using ROVER led to higher overall accuracy and lower correlation with ASR.

While the word-level agreement based on majority vote was relatively low, we found that the number of Turkers who correctly transcribed the word was strongly related to the probability of the word being recognized by ASR: more Turkers correctly transcribed the words that were recognized by ASR and conversely the ASR correctly recognized a higher percentage of words that were recognized by a larger number of Turkers. Thus while 84% of words recognized by all 5 Turkers were also recognized for ASR, ASR correctly recognized only 48% of words that were recognized by only one of five Turkers. In other words, there is evidence the same words may present difficulties for both ASR and human transcribers.

Finally, we explored whether some of the errors in both human and ASR transcriptions may be due to minor morphological discrepancies (e.g. "box" vs. "boxes") or incorrect transcription of function words such as "a" or "the". We found that stemming transcriptions indeed led to higher accuracy for both ASR and human transcribers. Restricting data to content words resulted in an increase in recall-based accuracy and agreement. Somewhat surprisingly, this slightly reduced the WER accuracy of both systems in comparison to the reference transcription.

There are a few limitations to our current study. In some cases the alignment of reference fragments with audio may not be matched precisely, introducing the potential for Turker transcription errors at the beginnings and ends of some fragments in our crowdsourced set. We aligned the ASR output to the exact audio fragments that were given to the humans in order to make for a fair comparison. Nonetheless, the alignment differences may have minor effect on the raw calculations of WER and recall.

There is also the question of how best to compare ASR systems with humans. A direct comparison is difficult to make, considering their vast differences in available resources and background knowledge. In this study, we compared humans who have annotated small fragments of a spoken response. In some sense, this may make the data more comparable because the human transcribers received only a few words of context and the ASR system utilizes a 4-gram language model. However, since ASR was run on full response, it also had access to the three words on each side of the fragment that were not available to human transcribers. This may impact the validity of our comparison.

In the future we plan to further investigate the nature of discrepancies between the crowdsourced human and automated speech recognition sets. Previous work has shown that there is a variety of features that contribute to reduced word accuracy, such as speaker characteristics, prosody, and disfluencies [16]. We will investigate these specific differences more in these data in future work.

4. REFERENCES

[1] Zhirong Wang, Tanja Schulz, and Alex Waibel, "Comparison of acoustic model adaptation techniques for non-native speech," in *Proceedings of the IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), 2003, pp. 540–543.

[2] Klaus Zechner, "What did they actually say? Agreement and disagreement among transcribers of non-native spontaneous speech responses in an english proficiency test," in *Proceedings of Speech and Language Technology in Education (SLaTE)*, 2009, pp. 3–6.

[3] Keelan Evanini, Derrick Higgins, and Klaus Zechner, "Using Amazon Mechanical Turk for transcription of non-native speech," in *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 2010, pp. 53–56, Association for Computational Linguistics.

[4] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] Christian Gaida, Patrick Lange, Rico Petrick, Patrick Proba, and David Suendermann-oeft, "Comparing Open-Source Speech Recognition Toolkits," Tech. Rep., DHBW Stuttgart, 2014.

[6] Shasha Xie, Keelan Evanini, and Klaus Zechner, "Exploring content features for automated speech scoring," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, 2012, pp. 103–111, Association for Computational Linguistics.

[7] Joost van Doremalen, Catia Cucchiarini, and Helmer Strik, "Optimizing automatic speech recognition for low-proficient non-native speakers," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–13, 2010.

[8] Alexei V. Ivanov, David Suendermann-Oeft, Vikram Ramanarayanan, Melissa Lopez, Keelan Evanini, and Jidong Tao, "Automated speech recognition technology for dialogue interaction with non-native interlocutors," in *Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, 2015, to appear.

[9] Jared Bernstein, Alistaire Van Moere, and Jian Cheng, "Validating automated speaking tests," *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[10] Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech & Language*, vol. 25, no. 2, pp. 282–306, Apr. 2011.

[11] Jidong Tao, Keelan Evanini, and Xinhao Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 294–299.

[12] Neeraj Deshmukh, Richard Jennings Duncan, Aravind Ganapathiraju, and Joseph Picone, "Benchmarking human performance for continuous speech recognition," in *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 2486–2489.

[13] William D Raymond, Mark Pitt, Keith Johnson, Elizabeth Hume, Matthew Makashay, Robin Dautricourt, and Craig Hilts, "An analysis of transcription consistency in spontaneous speech from the Buckeye corpus," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 1125–1128.

[14] Anastassia Loukina, Melissa Lopez, Keelan Evanini, David Suendermann-Oeft, and Klaus Zechner, "Expert and crowdsourced annotation of pronunciation errors for automatic scoring systems," in *Proceedings of Interspeech 2015*, 2015, to appear.

[15] Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 5270–5273.

[16] Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[17] Norihide Kitaoka, Daisuke Enami, and Seiichi Nakagawa, "Effect of acoustic and linguistic contexts on human and machine speech recognition," *Computer Speech & Language*, vol. 28, no. 3, pp. 769–787, 2014.

[18] Lei Chen and Su-Youn Yoon, "Detecting structural events for assessing non-native speech," in *Proceedings of the 6th workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2011, pp. 38–45.

[19] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," *The Journal of The Acoustical Society of America*, pp. 5687–5690, 2008.

[20] Jonathan G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 1997, pp. 347–354.