Spoken Text Difficulty Estimation Using Linguistic Features*

Su-Youn Yoon and **Yeonsuk Cho** and **Diane Napolitano**
Educational Testing Service
660 Rosedale Rd
Princeton, NJ, 08541, USA
syoon@ets.org

Abstract

We present an automated method for estimating the difficulty of spoken texts for use in generating items that assess non-native learners' listening proficiency. We collected information on the perceived difficulty of listening to various English monologue speech samples using a Likert-scale questionnaire distributed to 15 non-native English learners. We averaged the overall rating provided by three nonnative learners at different proficiency levels into an overall score of *listenability*. We then trained a multiple linear regression model with the listenability score as the dependent variable and features from both natural language and speech processing as the independent variables. Our method demonstrated a correlation of 0.76 with the listenability score, comparable to the agreement between the nonnative learners' ratings and the listenability score.

1 Introduction

Extensive research has been conducted on the prediction of difficulty of understanding written language based on linguistic features. This has resulted in various readability formulas, such as the Fry readability index and the Flesch-Kincaid formula, which is scaled to United States primary school grade levels. Compared to readability, research into listenability, the difficulty of comprehending spoken texts, has been somewhat limited. Given that spoken and written language share many linguistic features such as vocabulary and grammar, efforts were made to apply readability formula to the difficulty of spoken texts, rending promising results that the listenability of spoken texts could be reasonably predicted from readability formula without taking acoustic features of spoken language into account (Chall and Dial, 1948; Harwood, 1955; Rogers, 1962; Denbow, 1975; O'Keefe, 1971). However, linguistic features unique to spoken language such as speech rate, disfluency features, and phonological phenomena contribute to the processing difficulty of spoken texts as such linguistic features pose challenges at both perception (or parsing) and comprehension levels (Anderson, 2005). Research evidence indicated that ESL students performed better on listening comprehension tasks when the rate of speech was slowed and meaningful pauses were included (Blau, 1990; Brindley and Slatyer, 2002). Shohamy and Inbar (1991) observed that EFL students recalled most when the information was delivered in the form of a dialogue rather than a lecture or a news broadcast. The researchers attributed test takers poor performance on the latter two text types to a larger density of propositions, greater than that of the more orally oriented text type (p. 34). Furthermore, it is not difficult to imagine how other features unique to spoken language affect language processing. For example, prosodic features (e.g., stress, intonation) can aid listeners in focusing on key words and interpreting intended messages. Similarly, disfluency features (e.g., pause, repetitions) may provide the listener with more processing time and redundant in-

*We would like to thank to Yuan Wang for data collection, Kathy Sheehan for sharing text difficulty prediction system and insights, and Klaus Zechner, Larry Davis, Keelan Evanini, and anonymous reviewers for comments.

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 267–276,
San Diego, California, June 16, 2016.

Source	Length (sec.)	Number of passages	% in the total sample	Set A	Set B	Set C
English proficiency tests for business purpose	25 - 46	50	25	16	16	18
English proficiency tests for academic purpose	23 - 101	80	40	28	26	26
News	15 - 66	35	18	12	12	11
Interviews	30 - 93	35	18	11	12	12
Total		200	100	67	66	67

Table 1: Distribution of speech samples

formation (Cabrera and Martínez, 2001; Chiang and Dunkel, 1992). Dunkel et al. (1993) stated that a variety of linguistic features associated with spoken texts contribute to task difficulty on listening comprehension tests. Thus, for a valid evaluation of the difficulty of spoken texts, linguistic features relevant to spoken as well as written language should be carefully considered. However, none of the studies that we were aware of at the time of the current study had attempted to address this issue in developing an automated tool to evaluate the difficulty of spoken texts using linguistic features of both written and spoken language. Lack of an automated evaluation tool appropriate for spoken texts is evidenced in more recent studies that applied readability formula to evaluate the difficulty of spoken test directions (Cormier et al., 2011) and spoken police cautions (Eastwood and Snook, 2012).

Recently, Kotani et al. (2014) developed an automated method for predicting sentence-level listenability as part of an adaptive computer language learning and teaching system. One of the primary goals of the system is to provide learners with listening materials according to their second-language proficiency level. Thus, the listenability score assigned by this method is based on the learners' language proficiency and takes into account difficulties experienced across many levels of proficiency and the entire set of available materials. Their method used many features extracted from the learner's activities as well as new linguistic features that account for phonological characteristics of speech.

Our study explores a systematic way to measure the difficulty of spoken texts using natural language processing (NLP) technology. In contrast to Kotani et al. (2014)'s system for measuring sentence-level listenability, we predict a listenability score for a spoken text comprised of several sentences. We first gathered multiple language learners' perceptions of overall spoken text difficulty, which we operationalized as a criterion variable. We assumed that the linguistic difficulty of spoken texts relates to four major dimensions of spoken language: acoustic, lexical, grammatical, and discourse. As we identified linguistic features for the study, we attempted to represent each dimension in our model. Finally, we developed a multiple linear regression model to estimate our criterion variable using linguistic features. Thus, this study addresses the following questions:

- To what extent do non-native listeners agree with the difficulty of spoken texts?
- What linguistic features are strongly associated with the perceived difficulty of spoken texts?
- How accurately can an automated model based on linguistic features measuring four dimensions (Acoustic, Lexical, Grammatical, and Discourse) predict the perceived difficulty of spoken texts?

2 Data

2.1 Speech Samples

We used a total of 200 speech samples from two different types of sources: listening passages from an array of English proficiency tests for academic and business purposes, and samples from broadcast

news and interviews which are often used as listening practice materials for language learners. Table 1 shows the distribution of the 200 speech samples by source and by random partition into three distinct sets A, B, and C for the collection of human ratings. Each set includes a similar number of speech samples per source.

All speech samples were monologic speech and the length of speech samples was limited to a range of about 23 to 101 seconds. All samples were free from serious audio quality problems that would have obscured the contents. The samples from the English proficiency exams were spoken by native English speakers with high-quality pronunciation and typical Canadian, Australian, British, or American accents. The samples from the news clips were part of 1996 English Broadcast News Speech corpus described in Graff et al. (1997). We selected seven television news programs and extracted speech samples from the original anchors. The interview samples were excerpts from interview corpus described in Pitt et al. (2005). They were comprised of unconstrained conversational speech between native English speakers from the Midwestern United States and a variety of interviewers who, while speaking native- or near-native English, are from unknown origins. We only extracted a monologic portion from the interviewee.

2.2 Human Ratings

A questionnaire was designed to gather participants' perceptions of overall spoken text difficulty, operationalized as our criterion variable. The questionnaire is comprised of five Likert-type questions designed to be combined into a single composite score during analysis. Higher point responses indicated a lower degree of listening comprehension and a higher degree of text difficulty. The original questionnaire is as follows:

1. Which statement best represents the level of your understanding of the passage?

   5) Missed the main point
   4) Missed 2 key points
   3) Missed 1 key point
   2) Missed 1-2 minor points
   1) Understood everything

2. How would you rate your understanding of the passage?

   5) less than 60%
   4) 70%
   3) 80%
   2) 90%
   1) 100%

3. How much of the information in the passage can you remember?

   5) less than 60%
   4) 70%
   3) 80%
   2) 90%
   1) 100%

4. Estimate the number of words you missed or did not understand.

   5) more than 10 words
   4) 6-10 words
   3) 3-5 words
   2) 1-2 words
   1) none

5. The speech rate was

   5) fast
   4) somewhat fast
   3) neither fast nor slow
   2) somewhat slow
   1) slow

The first three questions were designed to estimate participants' overall comprehension of the spoken text. The fourth question, regarding the number of missed words, and the fifth question were designed to estimate the difficulty associated with the Vocabulary and Acoustic dimensions. We did not include separate questions related to the Grammar or Discourse dimensions.

Our aim was to recruit two non-native English speakers of beginner, intermediate, and advanced proficiency and have them rate each set of speech samples. We were able to recruit 15 non-native English leaner representing various native language groups including Chinese, Japanese, Korean, Thai,

and Turkish. Prior to evaluating the speech samples, participants were classified into one of the three proficiency levels based on the score they received on the TOEFL Practice Online(TPO). TPO is an online practice test which allows students to gain familiarity with the format of TOEFL, and we used a total score that was a composite score of four section scores: listening, reading, speaking, and writing. Each participant rated one set, approximately 67 speech samples. The participants were assigned to one of the three sets of speech samples with care taken to ensure that each set was evaluated by a group representing a wide range of proficiency levels. Table 2 summarizes the number of listeners at each proficiency level assigned to each set.

	Beginner	Intermediate	Advanced
Set A	2	1	2
Set B	1	1	3
Set C	2	1	2

Table 2: Distribution of non-native listeners

All participants attended a rating session which lasted about 1.5 hours. At the beginning of the rating session, the purpose and procedures of the study were explained to the participants. Since we were interested in the individual participants' personal perceptions of the difficulty of spoken texts, participants were told to use their own criteria and experience when answering the questionnaire. Participants worked independently and listened to each speech sample on the computer. The questionnaire was visible while the listening stimuli were playing; however, the ability to respond to it was disabled until the speech sample had been listened to in its entirety. After listening to each sample, the participants provided their judgments of spoken text difficulty by answering the questionnaire items. The speech samples within each set appear in random sequence to minimize the effect of the ordering of the samples on the ratings. Furthermore, to minimize the effect of listeners' fatigue on their ratings, they were given the option of pausing at any time during the session and resuming whenever ready.

Before creating a single composite score from five Likert-type questions, we first conducted correlation analysis using the entire dataset. We created all possible pairs among five Likert-type questions and calculated Pearson correlations between responses to paired questions. The responses to the first four questions were highly correlated with Pearson correlation coefficients ranging from 0.79 to 0.92. The correlations between Question 5 and the other four questions ranged between 0.49 and 0.61. The strong inter-correlations among different Likert-type questions suggested that these questions measured one aspect: the overall difficulty of spoken texts. Thus, instead of using each response from a different question separately, for each audio sample, we summed each individual participant's responses to the five questions. This resulted in a scale with a minimum score of 5 and maximum score of 25, where the higher score, the more difficult the text. Hereafter, we refer to an individual-listener's summed rating an aggregated score.

Since our system goal was to predict the averaged perceived difficulty of the speech samples across English learners at beginning, intermediate, and advanced levels, we used the average of three listeners' aggregated scores, one listener from each proficiency level. Going forward we will refer to this average rating as the *listenability* score. The mean and standard deviation of listenability scores were 17.3 and 4.6, respectively. We used this listenability score as our dependent variable during model building.

3 Method

3.1 Speech-Based Features

In order to capture the acoustic characteristics of speech samples, we used speech proficiency scoring system, an automated proficiency scoring system for spontaneous speech from non-native English speakers. speech proficiency scoring system creates an automated transcription using an automated speech recognition (ASR) system and does not require a manual transcription. However, in this study, when generating features for our listenability model, we used a forced alignment algorithm to align the audio sample against a manual transcription in order to avoid the influence of speech recognition errors. This created word- and phone-level transcriptions with time stamps. The system also computes pitch and power and calculates descriptive statistics

Dimension	Feature	Correlation with Average Human Difficulty Rating
Acoustic	Speaking rate in words per second	-0.42
	Number of silences per word	0.25
	Mean deviation of speech chunk	-0.30
	Mean distance between stressed syllables in seconds	0.25
	Variations in vowel durations	-0.30
Vocabulary	Number of noun collocations per clause	-0.27
	Type token ratio	0.33
	Normalized frequency of low frequency words	-0.49
	Average frequency of word types	-0.25
Grammar	Average words per sentence	-0.38
	Number of long sentences	-0.39
	Normalized number of sentences	0.45

Table 3: Correlation between linguistic features and listenability

such as the mean and standard deviation of both of these at the word and response level. Given the transcriptions with time stamps and descriptive features of pitch and power, speech proficiency scoring system produces around 100 features for automated proficiency scoring per input. However, because speech proficiency scoring system is designed to measure the non-native speaker's degree of language proficiency, and a large number of features assess distance between the non-native test takers' speech and the native speakers' norm. These features are not applicable to our data since all audio samples are from native speakers. After excluding these features, only 20 features proved to be useful for our study. The features were classified into three groups as follows:

- Fluency: Features in this group measure the degree of fluency in the speech flow; for example, speaking rate and the average length of speech chunk without disfluencies;
- Pause: Features in this group capture characteristics of silent pauses in speech; for example, the duration of silent pauses per word, the mean of silent pause duration, and the number of long silent pauses;
- Prosodic: Features in this group measure rhythm and durational variations in speech; for example, the mean distance between stressed syllables in syllables, and the relative frequency of stressed syllables.

3.2 Text-Based Features

Text-based features were generated on clean transcripts of the monologic speech using the text difficulty prediction system system. (Sheehan et al., 2014) The main goal of text difficulty prediction system is to provide an overall measure of text complexity, otherwise known as readability, an important subtask in the measurement of listenability. However, because of the differences between readability and listenability, only seven of the more than 200 linguistic features generated by text difficulty prediction system were selected for our model, four of which cover the Vocabulary construct and three of which cover our Grammar construct.

3.3 Model Building

Beginning with the full set of features generated by speech proficiency scoring system and text difficulty prediction system, we conducted a correlation analysis between these linguistic features and our human ratings. We used the entire dataset for correlation analysis due to the limited amount of available data. We selected our subset of features using the following procedure: first, we excluded a feature when its Pearson correlation coefficient with listenability scores was less than 0.25. In order to avoid collinearity in the listenability model, we excluded highly correlated features ($r \geq 0.8$). Next, the remaining features were classified into four groups (Acoustic, Vocabulary, Grammar, and Discourse) each containing the three features representing that

dimension with the highest correlations. The final, overall set of features used in our analysis was selected to maximize the coverage of all of the combined characteristics represented by the overall constructs. For instance, if two features showed a correlation larger than 0.80, a feature whose dimension was not well represented by other features was selected. This resulted in a set of 12 features as presented in Table 3. We did attempt to develop a Coherence dimension using two features (the frequency of content word overlap and the frequency of casual conjuncts), but both were found to have insignificant correlations with the listenability score and thus were excluded from the model.

Model-building and evaluation were performed using three-fold cross-validation. We randomly divided out data into three sets, two of which were combined for training with the remaining set used for testing. For each round, a multiple linear regression model was built using the average difficulty ratings of three non-native listeners, one at each proficiency level, as the dependent variables and the 12 features as independent variables.

4 Results

4.1 Agreement among non-native listeners

In this study, we estimated the difficulty of understanding spoken texts based on self-reported ratings via Likert-type questions, similar to the approach taken by Kotani et al. (2014). Likert-type questions are effective in collecting the participants' impression for the given item and are widely used in survey research but are highly susceptible. Participants may avoid selecting extreme response categories (central tendency bias) or may choose the "easy" category more often to inflate their listening comprehension level. These distortions may result in shrinkage of the listenability score's scale. In particular, the second bias may be more salient for participants at low proficiency levels and cause a skew toward higher listenability scores. In order to examine whether any participant was subject to such biases, we first analyzed the distribution of response categories per each participant. Approximately 335 responses were available per participant (67 audio samples, 5 questions per sample). All participants made use of every response category, and 10 out of

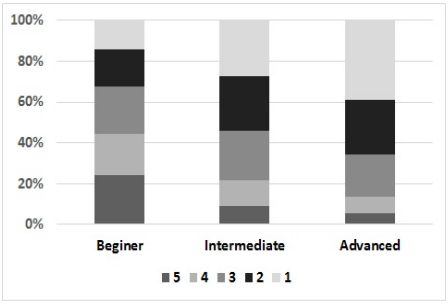

Figure 1: Distribution of Likert-type responses per proficiency group

15 participants used all categories at least 4% of the time. However, four participants rarely used certain response categories; two advanced learners and one intermediate learner used category "5" (most difficult) only 1%. On the contrary, one rater at the beginner level used category "1" (easiest) only for 1%. Due to the potential bias in these ratings, we tried to exclude them when selecting three listeners (one listeners per proficiency level) to use in calculating the listenability score; these advanced learners and this beginner learner were excluded, but the intermediate learner was included due to lack of an alternative learner at the same proficiency level.

Next, we examined the relationship between difficulty ratings and non-native listeners' proficiency levels. Figure 4.1 shows distribution of aggregated scores per proficiency group.

The aggregated score reflects the degree of comprehension by non-native listeners. The lowest response category indicated understanding of all words and possibly all main points, while the highest response category indicated that listeners failed to understand the main point, or they understood less than 60% of the contents. Beginners' scores were relatively evenly distributed; the proportion of response category "1" (easiest) was 14%, while the proportion of response category "5" (most difficult) was 24%. In regards to the high proportion of "5" responses by beginners, we would expect that, if there was a tendency on the part of the beginners to inflate their scores, the proportion of this category would be low. On the contrary, it was the most frequently selected category, demonstrating that the beginning listeners in this study did not seem to be inflating their ability to understand the spoken text.

Not surprisingly, as proficiency level increased, the listeners were more likely to judge the samples as easy, and the frequency of selecting categories representing difficulty decreased. The percentages of response category "5" selections were 24% for beginners, 9.1% for intermediate learners, and 5.3% for advanced learners.

Finally, we used Pearson correlation coefficients to assess the inter-rater agreement on the difficulty of spoken texts. The correlation analysis results between two listeners at the same proficiency level are summarized in second and third rows of Table 4. For the beginner group, the correlation coefficient for set B was unavailable due to the lack of a second listener. We also analyzed the agreement between all possible pairs of listeners across the different groups by calculating the Pearson correlation coefficient per pair and taking the average for each set (8 pairs for set A and C, 5 pairs for set B). The results are presented in the last row of Table 4.

Table 4 provides Pearson correlation coefficients.

Group	Proficiency Level	A	B	C	Mean
Within Group	Beginner	0.56	-	0.60	0.58
	Advanced	0.55	0.64	0.64	0.61
Cross-Group		0.61	0.58	0.60	0.60

Table 4: Pearson correlations among non-native listeners' ratings

The non-native listeners showed moderate agreement on the difficulty of our selection of spoken texts. Within the same group, the Pearson correlation coefficients ranged from 0.55 to 0.64, and the average was 0.58 for the beginner group and 0.61 for the advanced group. The average correlation across groups was also comparable to the within-group correlation values, although the range of the coefficients was wider, ranging from 0.51 to 0.7.

Next, we evaluated the reliability of the listenability scores (the average of three non-native listeners' ratings) based on the correlation with the second listener's ratings not used in the listenability scores. Compared to correlations between individual listeners' ratings (Pearson correlation coefficients of within-group condition), there were increases in the Pearson correlation coefficients. The Pearson correlation coefficient with the beginner group listener score was 0.65, and that with the advanced group listener score was 0.71; there was 0.07 increase in the beginner listener and 0.10 increase in the advanced listener, respectively. This improvement is expected since the listenability scores are averages of three scores and therefore a better estimate of the true score. We will use Pearson correlation coefficients of 0.65 and 0.71 as reference of human performance when comparing with machine performance.

4.2 Relationships Between Listenability Scores and Linguistic Features

We conducted a correlation analysis between our set of 12 features used in the model and the average listenability scores. A brief description, relevant dimension, and Pearson correlation coefficients with the listenability scores are presented in Table 3. Features in the Acoustic dimension were generated using speech proficiency scoring system based on both a audio file and its manual transcription. Features in both the Vocabulary and Grammar dimensions were generated using text difficulty prediction system and only made use of the transcription.

The features showed moderate correlation with the listenability scores, with coefficients ranging from 0.25 to 0.50 in absolute value. The best performing feature was the "normalized frequency of low frequency words" which measures vocabulary difficulty. It was followed by the "normalized number of sentences" which measures syntactic complexity and then the "speaking rate of spoken texts" from the Acoustic dimension.

4.3 Performance of the Automated System

Table 5 presents the agreement between ratings generated by our system and the human ratings. The model using both written and spoken features, "All", has a strong correlation with the averaged listenability score, with a Pearson correlation coefficient of 0.76. This result is comparable to the agreement between the average listenability score and those of the individual listeners (0.65 and 0.71). In order to evaluate the impact of different sets of features, we developed two models: a model based only on speech proficiency scoring system features (Acoustic dimension alone) and a model based only on text difficulty prediction system features (the Vocabulary

and Grammar dimensions). The performance of the model was promising, but there was a substantial drop in agreement: a decrease of approximately 0.1 in the Pearson correlation coefficient from the observed for the model with both written and spoken features. Overall, the results strongly suggest that the combination of acoustic-based features and text-based features can achieve a substantial improvement in predicting the difficulty of spoken texts over the limited linguistic features typically used in traditional readability formulas.

Feature Set	Correlation	Weighted Kappa
All	0.76	0.73
speech proficiency scoring system only	0.67	0.64
text difficulty prediction system only	0.65	0.63

Table 5: Correlation between automated scores and listenability scores based on human ratings

5 Discussion

Due to the limited amount of data available to us, the features used in the scoring models were selected using all of our data, including the evaluation partitions; this may result in an inflation of model performance. Additionally, we selected a subset of features based on correlations with listenability scores and expert knowledge (construct relevance) but we did not use an automated feature selection algorithm. In a future study, we will address this issue by collecting a larger amount of data and making separate, fixed training and evaluation partitions.

In this study, we used non-native listeners' impression-based ratings as our criterion value. We did not provide any training session prior to collecting these ratings which were based on individual participants' own perceptions of the difficulty. The individual raters had a moderate amount of agreement on the difficulty of the spoken texts, but for use in training our model, the reliability of listenability scores based on the average of three raters was substantially higher. However, impression-based ratings tend to be susceptible to raters' biases, so it is not always possible to get high-quality ratings. Ratings from non-native learners covering a wide range of proficiency levels is particularly difficult. Obtaining a high-quality criterion value has been a critical challenge in the development of many listenability systems. To address this issue, we explored automated methods that improve the quality of aggregated ratings. Snow et al. (2008) identified individual raters with biases and corrected them using small set of expert annotations. Ipeirotis et al. (2010) proposed a method using the EM algorithm without any gold data: they first initialize the correct rating for each task based on the majority vote outcome, then estimated the quality of each rater based on the confusion matrix between each individual rater's ratings and majority vote-based answers. Following that, they re-estimated correct answers based on the weighted vote using the rater's error rate. They repeated this process until it converged. Unfortunately we found that it was difficult to apply these methods to our study. Both methods required correct answers across all raters (either based on expert annotations or majority voting rules). In our case, the answers varied across proficiency levels since our questions were in regards to the degree of spoken text comprehension. In order to apply these methods, we would have needed to define a set of correct answers per proficiency level. In the future, instead of applying these automated methods exactly, we intend to develop a new criterion value based on an objective measure of a listener's comprehension. We will create a list of comprehension questions specific to each spoken text and estimate the difficulty based on the proportion of correct answers.

Originally, responses of individual Likert-type question are ordinal scale data. The numbers assigned to different response categories express a "greater than" relationship, and the intervals between two consequent points are not always identical. For instance, for the Likert-type question using five response categories ("strongly disagree", "disagree", "neither disagree nor agree", "agree", and "strongly agree"), the interval between "strongly agree" and "agree" may not be identical to the interval between "agree" and "neither disagree nor agree". Thus, some analyses applicable to interval data are not appropriate for Likert-type data. On the contrary, the Likert-scale data is comprised of a se-

ries of Likert-type questions addressing one aspect, and all questions are designed to create one single composite score. For this type of data, we can use descriptive analysis such as mean and standard deviation and linear regression models. In this study, five Likert-type questions were designed to measure one aspect, perceptions of overall spoken text difficulty, and, in fact, responses to different questions were strongly correlated. Based on this observation, we treated our data as a Likert-scale data and conducted various analysis applicable to the interval scale data.

Our method was initially designed to assist with the generation of listening items for language proficiency tests. Therefore, we focused on spoken texts frequently used on such tests, so, as a result, the range of text types investigated was narrow and quite homogenous. Interactive dialogues and discussions were not included in this study. Furthermore, although effort was made to include a variety of monologues by adding radio broadcasts to our data sample, a significant portion of the speech samples were recorded spoken texts that were designed for a specific purpose, that is, testing English language proficiency. It is possible that the language used in such texts is more contrived than that of monologues encountered in everyday life, particularly since they do not contain any background noise and were produced by speakers from a narrow set of English accents. That having been said, our method is applicable within this context, and predicting the difficulty of monologues produced by native speakers with good audio quality is its best usage.

6 Conclusion

This study investigated whether the difficulty of comprehending spoken texts, known as *listenability*, can be predicted using a certain set of linguistic features. We used existing natural language and speech processing techniques to propose a listenability estimation model. This study combined written and spoken text evaluation tools to extract features and build a multiple regression model that predicts human perceptions of difficulty on short monologues. The results showed that a combination of 12 such features addressing the Acoustic, Vocabulary, and Grammar dimensions achieved a correlation of 0.76 with human perceptions of spoken text difficulty.

References

John R Anderson. 2005. *Cognitive psychology and its implications*. Macmillan.

Eileen K Blau. 1990. The effect of syntax, speed, and pauses on listening comprehension. *TESOL quarterly*, 24(4):746–753.

Geoff Brindley and Helen Slatyer. 2002. Exploring task difficulty in esl listening assessment. *Language Testing*, 19(4):369–394.

Marcos Penate Cabrera and Plácido Bazo Martínez. 2001. The effects of repetition, comprehension checks, and gestures, on primary school children in an efl situation. *ELT journal*, 55(3):281–288.

Jeanne S Chall and Harold E Dial. 1948. Predicting listener understanding and interest in newscasts. *Educational Research Bulletin*, pages 141–168.

Chung Shing Chiang and Patricia Dunkel. 1992. The effect of speech modification, prior knowledge, and listening proficiency on efl lecture learning. *TESOL quarterly*, 26(2):345–374.

Damien C Cormier, Kevin S McGrew, and Jeffrey J Evans. 2011. Quantifying the degree of linguistic demand in spoken intelligence test directions. *Journal of Psychoeducational Assessment*, 29(6):515–533.

Carl Jon Denbow. 1975. Listenability and readability: An experimental investigation. *Journalism and Mass Communication Quarterly*, 52(2):285.

Patricia Dunkel, Grant Henning, and Craig Chaudron. 1993. The assessment of an l2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal*, 77(2):180–191.

Joseph Eastwood and Brent Snook. 2012. The effect of listenability factors on the comprehension of police cautions. *Law and human behavior*, 36(3):177.

David Graff, Zhibiao Wu, Robert MacIntyre, and Mark Liberman. 1997. The 1996 broadcast news speech and language-model corpus. In *Proceedings of the DARPA Workshop on Spoken Language technology*, pages 11–14.

Kenneth A Harwood. 1955. I. listenability and readability. *Communications Monographs*, 22(1):49–53.

Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM.

Katsunori Kotani, Shota Ueda, Takehiko Yoshimi, and Hiroaki Nanjo. 2014. A listenability measuring method for an adaptive computer-assisted language learning and teaching system. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*, pages 387–394.

M Timothy O'Keefe. 1971. The comparative listenability of shortwave broadcasts. *Journalism Quarterly*, 48(4):744–748.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

John R Rogers. 1962. A formula for predicting the comprehension level of material to be presented orally. *The journal of educational research*, 56(4):218–220.

Kathleen M. Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The textevaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, 115(2):184 – 209.

Elana Shohamy and Ofra Inbar. 1991. Validation of listening comprehension tests: The effect of text and question type. *Language testing*, 8(1):23–40.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.