An Evaluation of the Use of Text-Based Comprehensibility Measures on Online Spoken Language Learning Materials

Michael Gringo Angelo Bayona
Trinity College Dublin
Dublin, Ireland
bayonam@tcd.ie

Andrew Hines
University College Dublin
Dublin, Ireland
andrew.hines@ucd.ie

Emer Gilmartin
Trinity College Dublin
Dublin, Ireland
gilmare@tcd.ie

Elaine Uí Dhonnchadha
Trinity College Dublin
Dublin, Ireland
uidhonne@tcd.ie

***Abstract*— Automated assessment of online spoken materials for use in listening training can be of benefit to language learners. ICALL systems with this facility can help learners sift through collections of online content to find materials that suit their proficiency level and learning goals. Our pilot study assesses the extent to which readability measures can contribute to the listenability assessment of online spoken materials for language learning. Extending these text-based measures to assess listenability is convenient, but their use must be reassessed against spoken materials available online. We compare assessments by four readability measures with expert-assigned assessments of difficulty on online spoken language learning materials. We also evaluate the robustness of these measures against errors arising from automatically generating transcripts and their capability to predict human-based judgments of listenability. Our findings show that these comprehensibility measures can be combined and used to discriminate between materials of different levels of difficulty. We are also able to illustrate a way to conduct a fully automated approach to listenability assessment of spoken language learning materials via text-based measures. Our study also highlights the need for an approach that looks beyond the text, particularly for assessing higher-level materials and spoken materials of different genres.**

Keywords—listenability, assessment, language learning

I. Introduction

Second language learners need to practice their listening skills to become effective oral communicators in their target language. They can make use of authentic materials to be exposed to the naturalness of the target language and experience something close to real-life listening situations. It is recommended that learners have access to such materials as early as possible exactly for these mentioned reasons [1].

The Internet has made it easier for language learners to access authentic materials, which form a large part of what is available online. However, finding material that caters to a learner's evolving needs is neither easy nor straightforward. Language learners can be at different levels of proficiency and have various learning goals. These must be considered in the material selection. Learners can get listening suggestions from their teachers, but more self-directed or independent learners will want to source materials themselves. Choosing suitable material is vital as it can help learners overcome comprehension difficulties [2] and maintain motivation [3].

Text-based measures offer a convenient way of analyzing the listenability, or comprehensibility of spoken materials. Such approaches decouple the audio from the transcript and focus only on the material's content to identify the level of competency required of language learners who can use the material. Readability formulae such as Flesch's Reading Ease [4] and Dale-Chall Readability [5] formulae have been used as an objective way to assign proficiency level requirements on spoken learning materials [6, 7]. While multimodal approaches to listenability assessment that use text-, speech- and learner-based features [8, 9] already exist, purely text-based methods have been continually explored for listenability assessment [10, 11].

Regardless of the approach, automated means to listenability assessment can be a valuable tool for language learning. Intelligent computer-assisted language learning (ICALL) applications with this facility can help learners to efficiently navigate through the vast selection of materials that vary widely in content and delivery. ICALL applications can recommend spoken material suitable for every language learner based on their assessments of the listenability of the material and their perceived proficiency of the learner.

In this pilot study, we explored the extent by which the comprehensibility of online spoken materials can be reflected through an analysis of their transcripts. We looked at established readability metrics and assessed whether these measures can be used in evaluating the listenability of these materials found online. We also examined their robustness when using transcripts generated automatically. This paper details how well readability measures agree with human evaluations of listenability. It also demonstrates the maturity of artificial intelligence for the task by presenting an approach to fully automate the listenability assessment of materials.

II. Methodology

To examine how well text-based measures can be used for the listenability assessment of spoken materials, we evaluate listening activities from Voice of America (VOA) Learning English[1]. First, we study the extent to which these text-based measures transfer to the spoken domain and compare it with human evaluations of listenability. Second, we examine the robustness of these measures against transcripts generated automatically. Finally, we consider how well these measures predict human evaluations of listenability.

A. Data collection and preparation

The VOA website hosts a collection of materials directed at learners of English at various levels of proficiency. In this experiment we only consider news stories for our assessment so that the materials would fall under one general genre. Audio recordings of VOA news stories and their corresponding transcripts were downloaded from the website via a Python-based web crawler[2] that we implemented. A total of 1000 recordings were collected, sampled evenly across the three levels as defined by VOA content experts: beginner,

[1] https://learningenglish.voanews.com/

[2] Permission to download the materials were first secured from VOA

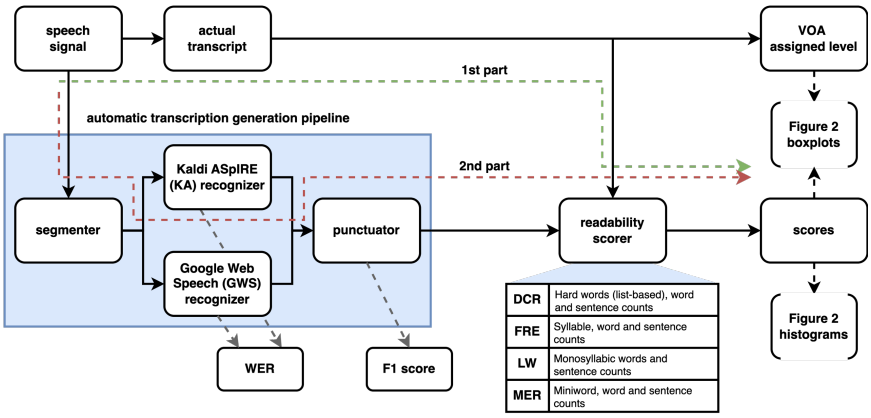

Fig. 1. Schematic diagram of the listenability assessment process

intermediate and advanced. Beginner-level materials are composed of video clips of news reporting and a segment defining an unfamiliar word found in the news. Intermediate- and advanced-level materials are audio recordings of news stories on different topics. Beginner-level materials are short, one-minute videos, while materials from other levels can be as long as ten minutes. Transcripts were normalized, which involved removing sections that were not present in the recording and spelling out numeric and alphanumeric tokens based on how they were spoken. This eased the accurate calculation of the readability scores and other performance metrics used in the experiment.

B. *Description of readability measures*

Four readability measures were used for this experiment and are listed in Fig. 1. Two of them are readability formulae that were previously used for listenability assessment research, namely the Flesch's Reading Ease (FRE) [4] and the Dale-Chall Readability (DCR) [5] formulae. The FRE formula considers word and sentence counts while the DCR formula focuses on the presence of easy and hard words, using a word list of easy words. The other two readability measures were included as they use at least one text-based feature that is different from FRE and DCR. First is the Lensear Write formula [12], which counts monosyllabic words in addition to sentences. Monosyllables were used by O'Hayre [12] to encourage the use of strong and active verbs which were assumed to be mostly monosyllabic. The other is the McAlpine EFLAW™ Readability (MER) formula [13] counts the number of miniwords, which are words of three letters or less, in addition to words and sentences. The count of miniwords was included by McAlpine [13] in her readability formula as miniwords were considered to increase confusion by nonnative readers of English. These are among many metrics that are widely used in the assessment of written texts.

The calculation of the readability scores is facilitated by various tools. Counts of syllables and monosyllabic words are generated using an algorithm based on the Maximal Onset Principle [14], and uses the CMU Pronouncing Dictionary[3] and a grapheme-to-phoneme system [15] to generate the phone sequence needed to determine the syllables counts. The computation of other text statistics is facilitated by the textstat [16], spaCy and NLTK libraries available in Python.

The readability scores are an estimate of the comprehensibility of a material. For FRE and LW, the lower the score, the more difficult the text is. For DCR and MER, lower scores correspond to easier materials. The scores are also mapped to readability levels. These levels correspond to a description of the target reader's proficiency level or the perceptual difficulty of the material. For DCR, the readability levels refer to U.S. grade levels [5], while the FRE and MER formulae map the levels to descriptions on how difficult the material will be for the learner [4, 13]. Meanwhile, the LW formula suggests that materials comprehensible by the average adult reader fall within a range of scores [12].

C. *Comparison of evaluations from readability measures and humans*

The audio recordings were first assessed based on their actual VOA-provided transcripts. This is illustrated in Fig. 1 and labeled "1st part". The process was carried out based on the processes of calculating the FRE and DCR scores [4, 5]. For short transcripts, defined as those with 300 words or less, the entire text was used for assessment. For long transcripts or texts with more than 300 words, three text samples were taken instead: one each from the beginning, middle and end of the transcript. Each sample is around 100 words, ensuring that samples do not begin or end in the middle of a sentence. The readability score for materials with long transcripts is then the average of the scores computed from the text samples.

The readability scores were then grouped based on the assigned VOA level of the material and boxplots were generated against the range of possible scores. This allows for the assessment of how well the VOA levels map to the scores derived from the metrics.

D. *Assessment of the robustness of readability measures*

We then evaluate the robustness of the four measures by looking at the effect of using automatically generated transcripts in the calculation of readability scores. We consider this since in many cases, transcripts of potential learning materials are unavailable, and it would be useful to have a system that can generate transcripts for assessment.

We use a pipeline to transcribe the audio recordings, labeled "2nd part" in Fig. 1. The pipeline has three blocks: a segmenter, an automatic speech recognizer (ASR), and a punctuator. Two aspects were considered: the effect of using the pipeline on the assessment of the material, and the consistency of the process when using different ASRs.

A silence-based segmenter [17] sits at the beginning of the pipeline to locate speech segments in the recording and send them to the ASR one at a time. This is necessary as the ASR

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

can sometimes fail to provide an output when the recording is too long. Two ASRs were used in deriving the transcript from the audio recording. The first is the ASpIRE Chain Model [18], implemented via the Kaldi Speech Recognition Toolkit. The second is the Google Web Speech API made available via the SpeechRecognition package [19] in Python. These ASRs were chosen for their functionality, availability, and their ability to handle different acoustic environments.

The ASRs determine the most likely word sequence that matches the recording but cannot identify where punctuations will be placed in the output text. Punctuation is critical in the calculation of the readability score since the formulae considered in this experiment include sentence counts. Thus, a punctuator [20] was added. The system uses a bidirectional recurrent neural network model with attention mechanism in placing the punctuation. This allows the punctuator to manage varying length contexts before and after a current position in text and identify relevant contexts.

The performance of the ASRs and the punctuator were evaluated against the VOA Learning English data previously collected. For the ASRs, the word error rate (WER) metric was used. It calculates the ratio of the sum of substitutions, insertions and deletions seen on the ASR output to the total number of words in the actual transcript. An F1 score was used to describe the punctuator's performance. This metric is calculated on the three most common punctuation marks found in the transcripts: period, comma, and question mark.

The readability scores were then computed from the transcripts generated by the pipeline using the same process described in subsection C of this section. Kernel density estimate (KDE) plots of these scores and those from the actual transcripts were plotted for analysis and comparison. These plots will allow us to easily identify any differences in the spread of the scores.

E. Predicting human-based listenability assessments

For the last part of the experiment, we assess how well the text-based measures predict human assessments of listenability. We examine the use of one text-based measure, as well as combinations of different measures in a classification task. Since the text-based formulae use different statistics in determining the comprehensibility of a material, we are interested in learning how well they work together in predicting the VOA level of a spoken material that is assigned by one or a group of language teaching experts. In addition, we also wanted to examine the effect of using the pipeline-generated transcripts in the ability of the combined readability scores to distinctly represent VOA data of various levels.

We consider a two-way and three-way classification task. Considering the two extreme VOA levels – beginner and advanced – as classes for the two-way task, we ran logistic regression with the readability scores as features. A repeated ten-fold cross-validation with the data was adopted. This approach was used as it could obtain a better estimate of the performance of each classifier [21]. We repeated the cross-validation 15 times and averaged the accuracies. We looked at the average accuracies as we explored different combinations and numbers of features to represent the spoken materials, as well as different transcript versions in generating the scores. The goal is not to optimize a system for listenability assessment but to assess the utility of combining different readability scores in predicting expert-assigned listenability expressed as VOA levels. For the three-way task, we implement a one-vs-rest logistic regression.

III. Results

A. Comparison of evaluations from readability measures and humans

Boxplots of the readability scores derived from the actual transcripts are shown in the first row of Fig. 2. There are three boxplots for each readability measure, one for each VOA level. The boxplots across all four measures have long whiskers and short boxes, which indicate that the scores for all VOA levels are concentrated on a narrow range.

The range of scores for the three VOA levels overlap as well. Except for LW, the interquartile ranges (IQRs) for intermediate- and advanced-level scores, which are represented by the boxes, map to almost the same span of values. The notch areas of the boxes for the two VOA levels overlap as well. This suggests that the median scores for the intermediate and advanced levels may not differ significantly. For LW, this same observation is seen for beginner- and intermediate-level scores.

Referring to the readability level mapping of the scores, the IQRs of the plots in Fig. 2 fit inside the span of one to three levels only. However, there are different interpretations of those levels. For DCR, the IQRs are within the range of 7 to 10. DCR maps the materials with scores in this range as suitable for 9th grade to college-level learners [5]. With FRE, the IQRs are within the scores 30 and 70, which corresponds to "standard" to "fairly difficult" [4]. For LW, the IQRs lie between 50 and 60, which is interpreted to be "too complicated for the average adult reader" [12]. Finally, the IQRs of MER scores are within 0 to 26, and interpreted as "very easy" to "quite easy to understand" [13]. This shows that the use of different sets of text-based features can lead to different assessments.

B. Assessment of the robustness of readability measures

We looked at the performance of the ASRs and the punctuator against the VOA Learning English data collected for this experiment. The Kaldi- and Google-based systems registered WERs of 20% and 15% respectively. Meanwhile, the punctuator obtained an F1 score of 57%.

We visualized the distributions of readability scores calculated using the different transcript versions. Also in Fig. 2, KDE plots of the scores from the actual transcripts (second row), and from the Kaldi-ASpIRE-based (KA) and Google Web Speech API-based (GWS) pipelines (third and fourth rows, respectively) are shown. Tick marks along the x-axis are based on the readability levels for each formula, except for the LW distributions since this formula does not have a detailed mapping to different readability levels.

Across all readability measures, the distributions of scores derived from both KA and GWS transcripts are noticeably different from the KDE plots of scores derived from the actual transcripts. The KDE plots for MER have the most noticeable change, where the score distributions shift towards higher readability scores. This means that the VOA materials are evaluated to be at a more difficult level using this readability measure. Other shifts in distribution are less prominent, but still noticeable. These show that the inaccuracies of both ASRs and the punctuator used can be significant enough to change the perceived comprehensibility of the VOA materials.

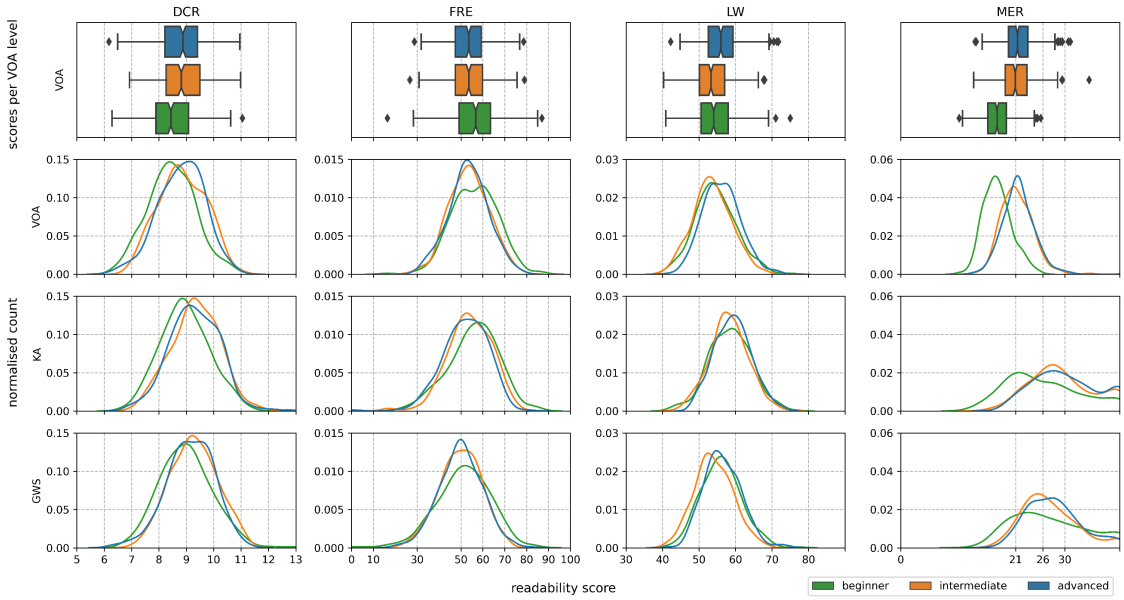

Fig. 2. Distribution of readability scores. *First row*: Boxplots of scores grouped by VOA human-labelled categories for different text-based measures using the actual transcripts. *Second to the fourth row*: Kernel density estimate plots of scores using actual transcripts (second row) and automatically generated transcripts (third and fourth rows). VOA denotes actual transcripts for calculating the scores, KA for transcripts from the Kaldi ASpIRE-based pipeline, and GWS for transcripts from the Google Web Speech API-based pipeline. Tick marks along the x-axis are based on the readability levels for each formula.

In addition, regardless of the transcript used, the distribution of readability scores according to the three VOA levels are overlapping. The most distinction seen among score distributions are with the MER KDE plots. Across all three setups, the distributions corresponding to the VOA beginner level have a peak location that is distinct from the other two distributions. The distributions of MER scores for the other two levels are concentrated on higher values and have greater overlap with each other.

C. Predicting human-based listenability assessments

The accuracies for predicting human-based listenability assessments using different readability score combinations are shown in Fig. 3 for the two-way classification task and Fig. 4 for the three-way classification task. The markers indicate the average accuracy of the repeated cross validation, and the bars show one standard deviation away from the average accuracy.

With readability scores derived from the actual transcripts as features for the two-way classification task, the accuracy increases as the number of features used for classification increases. For setups with only one feature used, average accuracies are at most 60%, apart from the use of the MER score, which sits above 75%. Meanwhile, setups using two or more features have average classification accuracies greater than 60% except for the setup using DCR and FRE scores. The highest average classification accuracy is almost 80%, for the setup that uses all four readability scores. However, we note that comparable average accuracies were recorded for setups that include the MER score in the set of features.

Meanwhile, using readability scores derived from GWS and KA transcripts yielded lower average accuracies for the two-way classification task compared to using scores from actual transcripts. The difference between the average accuracies is more apparent in setups using more than one readability score. Average accuracies for majority of KA setups are just above 60%, while average accuracies for GWS setups are still below 60%. While the use of more than one readability score result to an increase in average accuracy, this increase is not as significant as the one observed with setups using readability scores from actual transcripts. We also looked at which readability scores were used in classifiers that yielded high average classification accuracies. For both GWS and KA setups, there is no clear association between the readability measure used in the feature set and the average classification accuracy.

For the three-way classification task, lower average accuracies were recorded since it is more difficult compared to the two-way version. However, similar observations can be made. For using readability scores derived from actual transcripts, the use of two or more readability scores translates to a significant increase in average classification accuracy, with the maximum reached at 60% using all four readability scores. The MER score is also seen in setups that register a high average classification accuracy.

Meanwhile, average accuracies obtained for the three-way classification task with features derived from GWS and KA transcripts are lower. With a single feature, average accuracies obtained are comparable with setups using scores from actual transcripts. However, the difference in average accuracies is more apparent in majority of setups using more than one readability score. The maximum average accuracy for the KA setup is around 45%, while the maximum average accuracy for the GWS setup is below 45%. There is also no clear pattern seen between the readability measure used in the feature set and the average classification accuracy.

We note that we also looked at the learning curves for each of the setups used for the classification task. Shown in Fig. 5 is the plot of the learning curves for the setup that uses all four readability scores for the three-way classification task. This setup is characterized by both training and cross-validation scores settling at low accuracy values. Also, the two curves settle at values close to each other. This demonstrates high bias and that the data used is enough to derive the observations presented. This same case was observed in all the setups.

IV. Discussion

This study focuses on a small part of the automated spoken material assessment task. By looking at recordings containing news stories, we limit the scope to one general audio type characterized mostly by prepared texts narrated to listeners, with a few samples incorporating excerpts of interviews. The use of readability formulae allows us to assess the extent by which text statistics can be used to evaluate the listenability of a spoken material for language learning.

We first looked at how the readability levels associated with the different measures considered in this experiment map to the expert-assigned difficulty levels on the spoken materials. The overlapping ranges of readability scores among the three VOA levels show that the levels assigned by the VOA content creators cannot be satisfactorily mapped to readability levels using any of the formulae considered. The significant overlap of the boxes for VOA intermediate- and advanced-level materials suggests that from the point of view of any one readability formula used here, materials from these VOA levels can be rated as similar, suggesting that at these higher levels the differences in listenability cannot be characterized by measures based on text statistics alone.

We then examined the robustness of these readability measures against errors from automatically generating a transcript for assessment. The KDE plots of readability scores provided insights as to how the pipeline errors in transcript generation affect the assessments made by the different readability metrics. Among the four readability measures considered for this experiment, the MER formula is seen to be the least robust to errors from automatic transcript generation, as evidenced by the notable change in the KDE plots of scores derived from the KA and GWS transcripts. This also stresses the need for highly accurate systems for automation if we are to rely on only one readability measure for listenability assessment. Still, the pipeline presented illustrates the possibility of doing a fully automated approach to listenability assessment of spoken materials. While automated means of computing readability scores have been used in previous research [10], the pipeline allows a way to calculate the readability even when the material does not have a transcript.

Finally, we investigate how well readability measures can predict expert-assigned listenability. We see that readability scores can be a viable baseline system to distinguish between extreme levels of difficulty. Considering readability scores derived from actual transcripts, the binary classifier (beginner vs. advanced level) can already achieve an accuracy of more than 75% already, while it was more than 50% for the three-way classification task, both using MER scores. However, a significant drop in accuracy was observed when using scores derived from transcripts generated from both GWS and KA pipelines, reiterating the need for highly accurate systems.

There is improvement obtained with combining scores from different readability measures. The increase in accuracy is greater in the three-way classification task, suggesting the utility of combining different measures in more complicated classification tasks. However, the improvements are modest and the use of text statistics in general warrant further exploration. Still, referring to recent approaches that employ multimodal features [8, 9], though not directly comparable, suggests that the measures included in this experiment can be part of the features used for listenability assessment.

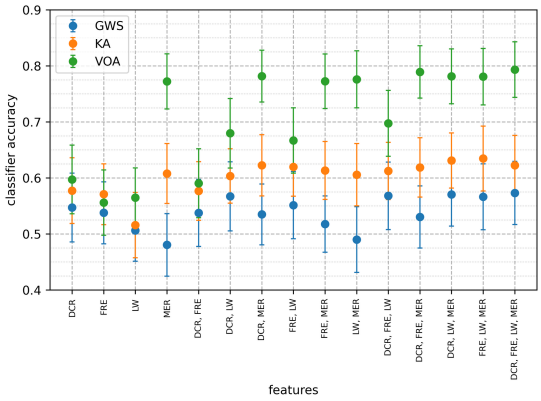

Fig. 3. Average accuracies and standard deviation for the two-way classification task

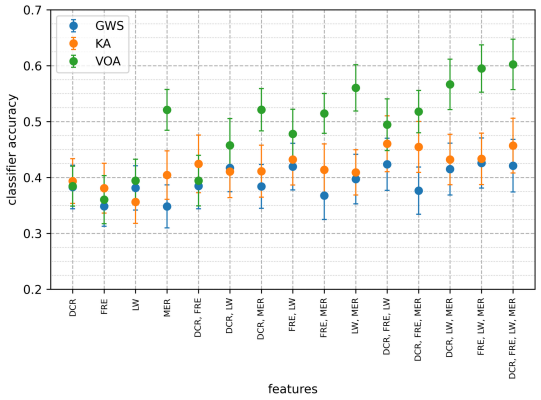

Fig. 4. Average accuracies and standard deviation for the three-way classification task

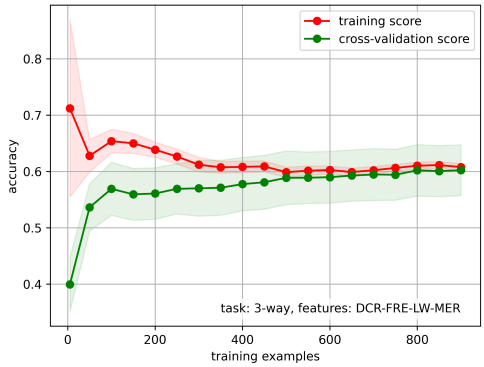

Fig. 5. Learning curves for the setup using all readability scores for the three-way classification task

In extending the scope of the experiment to include other online learning materials, we must consider the additional complexity this introduces to the task of listenability assessment. Spoken materials such as dialogues and talks differ from VOA news materials in that they can be less rehearsed, have different vocabulary used, and have different speech patterns. For recordings of more spontaneous nature, the pipeline must be able to handle hesitations or other disfluencies that may be present. These variations in speech patterns can also be a motivation to use speech-based features. Including other online materials also requires us to re-examine the consistency of rating the appropriateness of the materials according to the expected proficiency level of the learner. For example, given that more natural conversations can be more difficult for beginning language learners to comprehend, how will these materials be rated compared to the VOA materials?

Subsequent experiments must also consider learner input in developing the system. The current study only looked at levels assigned by the VOA content creators, however learner perceptions on the comprehensibility of a material may differ or be more varied and will thus be useful in distinguishing learning materials appropriate to different levels.

Finally, we note that the experiment was done on spoken materials written in English, and with measures specifically crafted for English. A study on cross-lingual readability assessment [22] show that for languages other than English, text statistics such as those used by the measures included here may not be the best predictors of comprehensibility. Thus, a separate study is warranted for other languages and perhaps focus on other text-based features.

V. Summary and Conclusion

In this paper we discussed the possible role of automated assessment of spoken material for language learning and explored the use of text-based measures towards developing such a system. We conducted a pilot study investigating the utility of readability metrics in assessing listenability in a variety of setups. We looked at how well the readability scores agree with expert evaluations, the robustness of the measures against the use of automatically generated transcripts, and their ability to predict expert evaluations of comprehensibility.

Automatically recognizing the words and providing punctuations both have a major influence on the readability scores as these text-based measures rely on features such as sentence length, number of sentences, and word list matching. Running readability on transcripts generated by our pipeline significantly decreased the classifier performance. This was a result of the ASR and the punctuator performance that make up our pipeline. When designing systems, we must be alert to the accuracy of the ASR and punctuator systems and sensitivity of the readability measures. At the very least, the pipeline presented here illustrates a way to conduct a text-based assessment approach even when the spoken material does not come with a transcript.

Still, our results show that measures based on text statistics can be a good baseline system to predict or replicate evaluations done by experts, at least for distinguishing between extreme difficulty levels. The MER readability measure stands out as better than the others on its own for ground truth labels (Fig. 3.) Mixing metrics in classifiers did not lead to significant improvements in classification accuracy. The combination of scores from all four readability measures considered in the experiment was best for two- and three-way classification task (Fig. 3 and Fig. 4) but not much better than using the MER measure on its own for two-way classification task and marginally better for three-way task.

Overall, this work highlights that readability is an important feature but cannot be relied upon in isolation for listenability. It is a weak feature label when computed via texts generated using ASR and a punctuator. However, it can still be a valuable contribution in a multimodal assessment when combined with other measures on the speech signal. Such approaches are the subject of our ongoing research.

Acknowledgment

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

References

[1] J. Field, "The Changing Face of Listening," in *Methodology in Language Teaching: An Anthology of Current Practice*, J. C. Richards and W. A. Renandya Eds., (Cambridge Professional Learning. Cambridge: Cambridge University Press, 2002, pp. 242-247.

[2] A. P. Gilakjani and N. B. Sabouri, "Learners' Listening Comprehension Difficulties in English Language Learning: A Literature Review," *English Language Teaching,* vol. 9, no. 6, pp. 123-133, May 06, 2016, doi: 10.5539/elt.v9n6p123.

[3] A. C.-S. Chang and S. Millett, "The effect of extensive listening on developing L2 listening fluency: some hard evidence," *ELT Journal,* vol. 68, no. 1, pp. 31-40, January, 2013, doi: 10.1093/elt/cct052.

[4] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology,* vol. 32, no. 3, pp. 221-233, June, 1948, doi: 10.1037/h0057532.

[5] E. Dale and J. S. Chall, "A Formula for Predicting Readability: Instructions," *Educational Research Bulletin,* vol. 27, no. 2, pp. 37-54, February 18, 1948, doi: 10.2307/1473669.

[6] J. Moldstad, "Readability Formulas and Film Grade-Placement," *Educational Technology Research and Development,* vol. 3, pp. 99-108, 1955, doi: 10.1007/BF02713358.

[7] M. T. O'Keefe, "The Comparative Listenability of Shortwave Broadcasts," *Journalism Quarterly,* vol. 48, no. 4, pp. 744-748, Winter, 1971, doi: 10.1177/107769907104800418.

[8] K. Kotani, S. Ueda, T. Yoshimi, and H. Nanjo, "A Listenability Measuring Method for an Adaptive Computer-assisted Language Learningand Teaching System," *28th Pacific Asia Conference on Language, Information and Computing (PACLIC),* pp. 387-394, 2014. [Online]. Available: https://aclanthology.org/Y14-1045.

[9] S.-Y. Yoon, Y. Cho, and D. Napolitano, "Spoken Text Difficulty Estimation Using Linguistic Features," *11th Workshop on Innovative Use of NLP for Building Educational Applications,* pp. 267-276, June 2016, doi: 10.18653/v1/W16-0531.

[10] T. M. Lowrey, "The Relation between Script Complexity and Commercial Memorability," *Journal of Advertising,* vol. 35, no. 3, pp. 7-15, 2006, doi: 10.2753/JOA0091-3367350301.

[11] M. S. Mirzaei and K. Meshgi, "Sentence Complexity as an Indicator of L2 Learner's Listening Difficulty," *EUROCALL 2020 Conference,* pp. 227-232, 14 December 2020. [Online]. Available: https://files.eric.ed.gov/fulltext/ED611114.pdf.

[12] J. O'Hayre, "Gobbledygook Has Gotta Go," Bureau of Land Management's Western Information Office, Denver, Colorado, 1966.

[13] R. McAlpine, *Global English for Global Business*. Wellington, New Zealand: CC Press, 2005.

[14] B. Hayes, "Syllables," in *Introductory Phonology*, 2009, ch. 13.

[15] K. Park and J. Kim. "g2PE." GitHub. https://github.com/Kyubyong/g2p (accessed November, 2021).

[16] A. Ward. "Textstat (Version 0.7.3)." GitHub. https://github.com/textstat/textstat (accessed November, 2021).

[17] J. Robert. "Pydub (Version 0.25.1)." GitHub. https://github.com/jiaaro/pydub (accessed November 2021).

[18] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASpIRE system: Robust LVCSR with TDNNS, iVector adaptation and RNN-LMS," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 13-17, 2015, pp. 539-546, doi: 10.1109/ASRU.2015.7404842.

[19] A. Zhang. "Speech Recognition (Version 3.8)." https://github.com/Uberi/speech_recognition (accessed November, 2021).

[20] O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," presented at the Interspeech 2016, San Francisco, USA, September 8-12, 2016.

[21] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics,* vol. 21, no. 15, pp. 3301-3307, 2005, doi: 10.1093/bioinformatics/bti499.

[22] I. Madrazo Azpiazu and M. S. Pera, "Is cross-lingual readability assessment possible?," *Journal of the Association for Information Science and Technology,* vol. 71, no. 6, pp. 644-656, 2020, doi: 10.1002/asi.24293.