EVALUATING ON-DEVICE ASR ON FIELD RECORDINGS FROM AN INTERACTIVE READING COMPANION

Anastassia Loukina[1], *Nitin Madnani*[1], *Beata Beigman Klebanov*[1], *Abhinav Misra*[1], *Georgi Angelov*[2], *Ognjen Todic*[3]

[1]Educational Testing Service, USA
[2]Astea Solutions, Bulgaria, [3]Keen Research, USA

ABSTRACT

Many applications designed to assess and improve oral reading fluency use automated speech recognition (ASR) to provide feedback to students, teachers, and parents. Most such applications rely on a distributed architecture with the speech recognition component located in the cloud. For interactive applications, this approach requires a reliable Internet connection that may not always be available. We investigate whether on-device ASR can be used for a virtual reading companion using recordings obtained from children both in a controlled environment and in the field. Our limited evaluation makes us cautiously optimistic about the feasibility of using on-device ASR for our application.

Index Terms— human-computer interaction, embedded recognizers, reading fluency assessment, child speech

1. INTRODUCTION

According to a 2015 report, 31% of U.S. 4th graders read below the Basic level.[1] In a typical upper-elementary curriculum, students are assumed to be sufficiently fluent independent readers to handle reading assignments in the science and social science classes. Thus, students who struggle to read start falling behind not only in English but across the curriculum. Moving from *learning to read* to *reading to learn* is one of the critical junctures in literacy development; it is important to create scalable methods for supporting struggling readers in this process.

Speech technologies offer rich opportunities for supporting literacy development. In previous work, we describe a virtual reading companion which uses the recording of an expert adult reader as a reading partner that takes turns reading a book with the child and employs speech processing technologies to process child's speech [1, 2]. During the narrator's turns, the child can follow along on the screen or just listen. When it is the child's turn to read, an automated speech analysis system captures and processes the child's oral reading in order to adjust the system's behavior (the narrator might read more to a weaker reader), provide feedback, or track improvement in reading skill. The application is intended to be used flexibly in both in-school and out-of-school scenarios, such as after-school programs, summer camps, or at home.

While automated systems have successfully been used for assessment of oral reading on short passages [3, 4], they primarily relied on distributed architecture with speech recognition components located in the cloud (although see [5]). In the context of our application, a reliable network connection may not always be available. According to the 2017 EdTech Outlook Report [6], while 94% of schools are connected to the Internet, only 22% of them have enough bandwidth to handle the streaming demands of media-rich applications. Furthermore, individual users might rely on cellular data plans and transmitting large amount of audio data might be financially prohibitive.

In addition to addressing the bandwidth limitation, [7] list other advantages of using on-device ASR. These include a simpler architecture on the server-side, a lower energy footprint, mitigated privacy concerns, enhanced options for voice activity detection, and bandwidth availability for other tasks that require network communication.

In this paper, we evaluate whether we can obtain reliable estimates of student reading accuracy from the audio recordings from our application using an on-device ASR system. An additional practical consideration is that the on-device ASR needs to be integrated into the application for usability testing long before sufficiently many children read the whole book to generate enough in-domain data for training the system. Therefore, we evaluate whether it is possible to achieve reasonable performance with an ASR system trained entirely on *pre-existing* data external to our application.

To assess the effect of data quality, we evaluate on several corpora: (1) high-quality recordings of adult speech; (2) recordings of fluent children collected in a quiet office; and (3) recordings collected under conditions more similar to the envisioned use — with struggling or moderately proficient readers in after-school programs.

[1] https://www.nationsreportcard.gov/reading_math_2015/#reading/acl?grade=4

2. RELATED WORK

Many existing commercial and research applications use ASR and other speech processing technologies to assess oral reading fluency and assist with its development. Reviews of earlier systems can be found in [8, 9, 4] among many others; [10] provide an overview of some of more recent developments in the area of technology-based literacy instruction. VersaReader [4] and Project LISTEN [3] are two of the most mature systems in this area. While their findings show that this approach is very effective and results in measurements that show high agreement with those assigned by human raters, the majority of such applications rely on a client-server architecture with the recording sent to a server for recognition and processing.

The feasibility of on-device ASR has been demonstrated in several papers using both Gaussian Mixture based models [11, 12, 13] and DNN-based models [14, 15, 16]. To our knowledge, the only reading application that uses on-device ASR is the self-administered app *Moby.Read* [5]. They report that the system achieves high agreement in terms of words correct per minute but do not provide any estimates of ASR performance. The app asks students to read a word-list and four short passages which means that the total size of the vocabulary/language model could be relatively small. In our case, the students read much longer texts. The challenges of extended reading stem not just from the duration of the audio signal to be processed. While it might be possible to design a rotating schedule for users or turn the noisy air-conditioner off during a one-time test that takes only a few minutes per student, the context of extended reading sessions over a period of several weeks makes consistently controlling the environment very difficult. Thus, there is a high likelihood of background noise, e.g. other children possibly reading the same text aloud at the same time.

3. DESCRIPTION OF THE DATA

We evaluated our on-device ASR system on three corpora. In all cases, the speakers read excerpts from the same book but the recording conditions and speaker characteristics varied across the three corpora. We selected Harry Potter and the Sorcerer's Stone (HP1) by J. K. Rowling as the book to be read. For the adult reader, we use the recorded narration by Jim Dale [17]. All participants in the two user studies (§3.2 and §3.3) were recruited following well-established IRB procedures.

3.1. Narrator corpus

We extracted 61 overlapping passages from the first chapter of the audiobook, which varied in length between 151 and 436 words. These were the same passages as those in the field data described in §3.3. The extracts were recognized automatically and the ASR hypothesis was compared to the original book.

We use this corpus as a sanity check to make sure that no errors were introduced during language model training and ASR configuration, and that there are no consistent data-independent patterns. The ASR system should have very low WER on this data since our system uses LibriSpeech models (see §4.1) that are a good fit for the male narrator whose reading is very well enunciated and free of disfluencies or any off-task speech; these professional recordings are also free of background noise.

3.2. Pilot corpus

The second corpus contains recordings that were collected early in the project with a goal of evaluating various system components. To that end, the task was not interactive reading but just plain reading of a few passages [1, 2]. It includes 63 recordings (2.2 hours) of 22 children reading 3 texts from HP1. At the time of the recording (April 2017), all children attended grades 2-4 (6-8 children per grade, 12 girls and 10 boys). All children read three passages from Chapter 1 (246, 226 and 306 words). The texts were presented on a laptop screen and captured via a Cyber Acoustics Pro Grade Stereo Headset microphone. To simulate classroom conditions, the recordings were conducted in an office with 2-3 children reading simultaneously.

While the age of the children matched that of the target population, the children were selected via a convenience sample and were very fluent and accurate readers in comparison to their peers. The experiment was set up in a way that children could not proceed to the next text until 3 minutes from the start of the recording. For this evaluation, we segmented the audio to only use the portion where the child read the text.

This corpus allows us to evaluate the system performance on children where we expect a mismatch in terms of acoustic models under otherwise favorable circumstances: fluent readers and reasonably quiet recording conditions.

3.3. Field corpus

The data for this study was collected in Summer-Fall 2017 using a research prototype of the actual interactive reading system. The first few chapters of the book were split into passages of varying length. The system displayed the book and alternated between playing the recorded narration for some passages and prompting the user to read aloud other passages. The system was deployed on laptops and the reading aloud was captured via the same headsets as for Pilot corpus.

The data collection was conducted at several after-school programs and lasted 5 days at each site. 36 children participated in the study (23 boys, 13 girls). Average age at the time of the recording was 9;2 (min 7;0, max 12;2).

The first session was used to administer various tests. During the remaining four days, the children interacted with

the program for 20 minutes a day. The total number of turns per session varied across children. Furthermore, not all children completed all four sessions.

We excluded recordings shorter than 30s and another 22 with off-task speech in the middle of the reading. The final corpus consists of 395 recorded turns or 14.3 hours of recordings.

4. ASR EVALUATION

4.1. Description of on-device ASR

We considered various options for on-device ASR, either open-sourced or commercial systems offering free trials. Many of the ASR APIs designed for use on portable devices such as the iOS & Android ones, rely on cloud-based processing and thus did not meet our requirement for offline use. Only two system met this requirement: *PocketSphinx* [11] and *KeenASR* which uses *Kaldi* [18]. We selected Kaldi-based *KeenASR* for this first round of experiments since we used Kaldi in previous experiments with server-side ASR [2].

We used the freely available trial version of the *KeenASR* SDK for iOS[2] and created a simple app that uses the SDK to stream an audio file through it and produce the 1-best ASR hypothesis. The app was written in Swift, cross-compiled using Apple XCode v9.2 on a 15-inch MacBook Pro, and then run on a 10.5 inch iPad Pro. Using the app, we captured the 1-best ASR hypotheses generated for each file in our corpora as well as its processing time. The decoding graph was compiled on the iPad Pro once — *before* any files were processed — and, therefore, graph compilation is excluded from our timing estimates.

We used acoustic models supplied with the trial version of *KeenASR*. These are publicly available *librispeech-nnet2-en-us* models trained on the LibriSpeech corpus [19]. We used custom language model trained on the book text following the same procedure as described in [2]. Separate models were trained for each chapter. As an example, the language model (ARPA) for chapter 1 contained 1,172 unigrams, 3,590 bigrams and 4,429 trigrams. We also added phonemic transcriptions for all OOV words to the lexicon.

4.2. Identification of off-task speech

The recordings in Pilot and Field corpora contained a certain amount of off-task speech. Furthermore, as showed in [2], background speech during silences can often be picked by ASR and recognized as part of the child's response - a phenomenon that was termed 'ghost' off-task speech.

Given our constrained language models, it is not meaningful to compute performance estimates on these parts of the recordings. To identify any instances of real or 'ghost' off-task speech either before or after the on-task reading, we used the algorithm described in [2, 20] where the ASR hypothesis is aligned back to the expected text in order to identify the first and last word in the hypotheses.

For reference transcriptions, the transcribers were asked to mark any instances of off-task speech. We used these annotations to filter out only the relevant portions of the transcript.

4.3. Evaluation metrics

We report three evaluation metrics.

First, for comparison with other studies, we report *ASR Word Error Rate* (WER), the standard measure of ASR performance, computed by comparing ASR hypothesis with human transcriptions provided by a professional transcription agency. We used NIST procedure to compute WER.

The second evaluation focuses on whether the ASR performance is sufficient to obtain reliable estimates of student reading skills, the main figure of interest for our application. Reading accuracy is computed as the total number of words that have been read correctly divided by the total number of words in the prompt. This measure is sensitive to deletions and substitutions but ignores insertions. We compute student accuracy twice, first using human transcriptions and then using ASR hypothesis. Since ASR-based estimates might be consistently lower or higher than those based on human transcription, in this study we focus on whether there is a linear relationship between the two estimates and report the *correlation*.

Finally, we also consider *processing time* (xRT): while this parameter is not directly related to measuring reading accuracy, it has important practical implications for any modern application. We compute decoding speed as processing time divided by the duration of the audio time.

5. RESULTS

5.1. ASR Word Error Rate

The recognizer performance on the Narrator corpus was very close to human transcribers with a mean WER of 0.3% (WER ranged from 0% to 3.6%). Thus the system performed as expected when evaluated on quiet recordings from an accurate adult reader.

For the Pilot corpus, the recognizer failed to produce a hypothesis for 1 out of 63 responses, probably because during the reading the child moved very close to the microphone leading to clipping. The average ASR WER for on-task part of the recording was 9.3% (WER varied from 2.9% to 25%). This is an encouraging result given the mis-match between the adult acoustic model and these young readers. In fact, the performance was comparable to that of an out-of-the box server-based system with the acoustic model optimized for

[2] https://keenresearch.com/keenasr-docs/keenasr-ios-introduction.html

Table 1. Distribution of student reading accuracy (%) for Pilot and Field corpora, and correlations (Spearman's ρ and Pearson's r) between transcription-based and ASR-based estimates of student reading accuracy at passage, session, and speaker level. Note that for the Pilot corpus, session-level and speaker-level estimates are the same since all recordings were collected in one day.

Corpus	Level	N	Accuracy (%)							Correlation	
			min.	25%	50%	75%	max.	mean	std.	ρ	r
Pilot	Passage	62	87.5	96.8	98.2	99.3	100	97.7	2.3	.46	.51
Pilot	Session	22	93.9	96.6	98.1	98.8	99.5	97.7	1.6	.74	.72
Field	Passage	304	0	45.2	95.0	98.3	100	73.0	36.3	.69	.83
Field	Session	99	0	36.1	91.2	97.5	99.8	67.7	36.7	.76	.82
Field	Speaker	35[3]	0	43.9	72.4	95.9	99.6	64.3	35.7	.85	.92

young but non-native speakers. The latter server-based system achieved a WER of 10% [2]. The performance is also in line with other state of the art systems: [21] reported a WER of 7% for read speech from young non-native English language learners; reviewing child ASR, [22] cite a WER of 8-12% for different systems.

Not all recordings in the Field corpus could be processed since the trial version of the KeenASR SDK limits file duration to 200s. Out of the 395 recordings we selected for analysis, 343 recordings were below this limit. Out of these 343 recordings, the ASR produced no hypothesis for 28 files (8%). For another 11 files (3%), the hypothesis consisted of 1 word. Further analysis showed that for 3 of these 39 (11 + 28) files, human transcriptions were also empty. For the remaining 36 files, ASR failure was likely caused by audio quality issues. Therefore, our analysis is done on 304 recordings (9.5 hours) where we were able to obtain a hypothesis. We used the same procedure as before to identify on-task speech and computed the WER between automatically identified on-task speech and parts of transcription marked as on-task. The distribution of WER for on-task speech was very skewed and varied from 1.8% to substantially above 100% [4] with a median value at 33.4%, substantially larger than for the Pilot corpus. To make sure that the WER results were not an artifact of the on-task detection algorithm, we also computed WER using the full transcriptions and hypotheses. This led to an increase in median WER to 36.1%. This is consistent with our expectation that given the constrained language model, ASR performance on off-task speech is likely to have high WER.

5.2. Decoding time

The decoding time for all corpora was consistently faster than the duration of the recording. The decoding time was the shortest and most consistent for the narrator with average xRT = 0.17 (SD = 0.009, max 0.18). For the Pilot corpus, the average xRT was 0.23. The decoding was the slowest for the recordings in the Field corpus with average xRT=0.5. The highest xRT across the three corpora was 1.45.

5.3. Reliability of reading accuracy estimates

We considered several accuracy estimates: (1) accuracy estimates for individual passages; (2) accuracy estimate for all passages read during a single session; and (3) accuracy estimate for all passages read by a given speaker during the entire data collection. Note that for the Pilot corpus, all recordings were collected in one day and, therefore, (2) and (3) are the same.

Table 1 shows the distribution of reading accuracy values computed based on the transcription and the correlations between the transcription-based and ASR-based accuracy estimates. We report both parametric (Pearson's r) and non-parametric correlations (Spearman's ρ) because in some cases the distributions are very skewed.

All estimates were consistently less reliable for the Pilot corpus than for the Field corpus. This may appear surprising given that the WER on Pilot corpus was much lower than for the Field corpus. Table 1 shows that the variation in student accuracy for the Field corpus is also much larger than for the Pilot corpus where all speakers were very accurate readers. ASR performance is clearly insufficient to make a fine-grained distinction between the highly proficient readers in the Pilot corpus. It is, however, useful in making coarser distinction between the greater range of abilities in the Field corpus.

We also observe that the accuracy estimates get more precise as several passages from the same speaker are aggregated to obtain a single estimate. Our ASR-based speaker-level estimates for the Field corpus are fairly precise with r=0.9 (ρ=0.84).

6. DISCUSSION

Our first observation is that on-device ASR is certainly a viable option for applications such as ours: its performance on high-quality recordings (Pilot corpus) is consistent with per-

[3] All recordings from one of the female speakers were excluded as a result of restriction on file duration.

[4] For our data, WER values $> 100\%$ usually indicate that the ASR hypothesis was longer than the human transcription.

formance reported for server-based ASR systems, even with mismatched acoustic models and in the absence of any adaptation for the data. The decoding speed is also generally quite fast with an average xRT of 0.5 for the Field corpus. We note, however, that the recordings used in this study were not collected using actual mobile devices which could lead to different performance estimates.

We also observed an increase in ASR WER when the system was evaluated on recordings in the Field corpus. Since these recordings were meant to be collected under authentic usage conditions, we only provided general guidelines to the sites (e.g. "try to place children as far from each other as possible" or "try to make sure the AC is off") but did *not* discard any sites even if it was clear that the acoustic conditions were less than perfect. As a result, many recordings in this corpus contained noise such as the speech of other students or from mechanical equipment.

To separate the effect of background noise, we rated the quality of all recordings on a scale of 1 to 3, where '1' corresponds to "I can barely understand hear or identify the main speaker because of noise", and '3' corresponds to 'The main speaker is clearly audible". Two annotators took part in the annotation process with about 50% of recordings annotated by both annotators. The inter-rater agreement for double-annotated recordings was quadratically weighted κ=0.70. Where available, we used the average of the ratings between the two annotators. We classified recordings with an average score of 1 as 'bad' (29.3%) and those with 2.5 or higher as 'good' (30.9%). The remaining 39.8% were considered 'noisy'.

Transcription-based estimates of reading accuracy were generally low for 'bad' recordings (median 21% vs. 95% for the whole corpus), partially due to the fact that the human transcriptions were not reliable because of background noise. Background noise was also a major factor in ASR performance: median WER was larger than 100% for 'bad' recordings, 28.5% for 'noisy' recordings, and 14.6% for 'good' recordings.

We observed that even for 'good' recordings, the WER was higher than for the Pilot corpus. We further reviewed some of the responses with particularly high WER and found that, in such cases, the student's reading deviated from the text in a non-trivial way: for example, the student switched between reading aloud and silent reading resulting in the recordings of only some parts of the text with long pauses between the utterances. During the pauses, the microphone picked up audio signals from the other children reading in the background and this was, in turn, recognized by the ASR system.

These problems are, of course, not restricted to the use of on-device ASR considered in this paper. For future work, we envision a number of technical approaches that could alleviate these problems, e.g. using noise-robust acoustic models, using other signals from the device like the front-facing camera to help with Voice Activity Detection, using more complex algorithms for the detection of on-task speech, and leveraging multiple microphones on the device for directional audio capture. Generally speaking, it is evident that a non-negligible percentage of recordings in authentic classroom conditions would *not* be suitable for obtaining reliable estimates of child's reading accuracy at the passage level.

At the same time, despite background noise and other unusual properties of the individual passages, our aggregated estimates of student accuracy remain very high. This suggests that noise and various other problems that adversely impact audio quality are distributed reasonably randomly across locations, sessions, and individuals; namely, it was not the case that specific locations consistently yielded low quality data. This is an encouraging result in terms of the general feasibility of using ASR-based technology for this application.

In this paper, we focused on a limited use case for the interactive reading app: providing a global estimate of the accuracy rather than fine-grained feedback, e.g., miscue analysis. Such feedback would require reliable measurements at the passage level and, therefore, additional filtering to exclude responses where such measurements are not likely to be reliable.

7. CONCLUSION

We evaluated an off-the-shelf, on-device ASR toolkit on field data collected from an interactive reading application under realistic classroom conditions. The results of our analysis make us cautiously optimistic about the feasibility of using on-device ASR for such an application. Using our ASR set-up as an example, we also showed that while background noise and off-task behavior lead to deterioration in performance in comparison with clean recordings, aggregating measurements across multiple responses from the same students can likely mitigate the issue, at least for coarse-grained accuracy estimates. For more fine-grained feedback at the passage level, additional filtering modules will be required at the very least.

8. ACKNOWLEDGEMENTS

We thank Blair Lehman for leading the field study; René Lawless for organizing data collections; John Sabatini and Tenaha O'Reilly for their advice on assessment of oral reading; Jason Bonthron, Tom Florek, Pavan Pillarisetti, and Nathan Lederer for building the data collection software; Zydrune Mladineo and Laura McCaulla for the annotations; Binod Gyawali for processing and preparing the e-book. We also thank our many colleagues at ETS for proctoring different sessions, useful discussion and comments. Finally, we thank all the participants and site administrators for allowing us to conduct the study.

9. REFERENCES

[1] Beata Beigman Klebanov, Anastassia Loukina, John Sabatini, and Tenaha O'Reilly, "Continuous fluency tracking and the challenges of varying text complexity," in *Procceings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, Copenhagen, Denmark., 2017, pp. 22–32, Association for Computational Linguistics.

[2] Anastassia Loukina, Beata Beigman Klebanov, Patrick Lange, Binod Gyawali, and Yao Qian, "Developing speech processing technologies for shared book reading with a computer," in *WOCCI 2017: 6th International Workshop on Child Computer Interaction*, ISCA, nov 2017, number November, pp. 46–51, ISCA.

[3] Jack Mostow, "Why and how our automated reading tutor listens," *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 43–52, 2012.

[4] Jennifer Balogh, Jared Bernstein, Jian Cheng, Alistair Van Moere, Brent Townshend, and Masanori Suzuki, "Validation of automated scoring of oral reading," *Educational and Psychological Measurement*, vol. 72, no. 3, pp. 435–452, 2012.

[5] Jared Bernstein, Jian Cheng, Jennifer Balogh, and Elizabeth Rosenfeld, "Studies of a Self-Administered Oral Reading Assessment," in *Proceedings of SLaTE 2017*, Olov Engwall, Jose Lopes, and Iolanda Leite, Eds., Stockholm, 2017, pp. 180–184, KTH Royal Institute of Technology.

[6] Chian Gong and Jennifer Carolan, "2017 EdTech outlook," https://www.edsurge.com/news/2017-11-17-spotting-the-2017-trends-that-fuel-edtech-innovation-and-investments.

[7] Alexei V Ivanov, Patrick L Lange, David Suendermann-oeft, Vikram Ramanarayana, Yao Qian, Zhou Yu, and Jidong Tao, "Speed vs . Accuracy : Designing an Optimal ASR System for Spontaneous Non-Native Speech in a Real-Time Application," in *Proceedings of IWSDS*, 2016, pp. 1–12.

[8] Maxine Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[9] Klaus Zechner, John Sabatini, and Lei Chen, "Automatic Scoring of Children's Read-Aloud Text Passages and Word Lists," *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 10–18, 2009.

[10] Scott A. Crossley and Danielle McNamara, "Educational Technologies and Literacy Development," in *Adaptive Educational technologies for literacy instruction*, Scott A. Crossley and Danielle Mcnamara, Eds., pp. 1–12. Routledge, New York, 2016.

[11] D. Huggins-Daines, Mohit Kumar, Arthur Chan, A.W. Black, Mosur Ravishankar, and A.I. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. 2006, vol. 1, pp. I–185–I–188, IEEE.

[12] Rohit Prasad, Prem Natarajan, David Stallard, Shirin Saleem, Shankar Ananthakrishnan, Stavros Tsakalidis, Chia lin Kao, Fred Choi, Ralf Meermeier, Mark Rawls, Jacob Devlin, Kriste Krstovski, and Aaron Challenner, "BBN TransTalk: Robust multilingual two-way speech-to-speech translation for mobile platforms," *Computer Speech & Language*, vol. 27, no. 2, pp. 475 – 491, 2013, Special Issue on Speech-speech translation.

[13] C Gaida, R Petrick, and D Suendermann-Oeft, "Kaldi Goes Android," in *Paper presented at Speech Ventures, Special Event at Interspeech 2016*, 2016.

[14] Xin Lei, Andrew Senior, Alexander Gruenstein, and Jeffrey Sorensen, "Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices," *Interspeech 2013*, , no. August, pp. 662–665, 2013.

[15] Andrew Senior and Xin Lei, "Fine context, low-rank, softplus deep neural networks for mobile speech recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 7644–7648, 2014.

[16] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Hasim Sak, Alexander Gruenstein, Francoise Beaufays, and Carolina Parada, "Personalized speech recognition on mobile devices," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 5955–5959, IEEE.

[17] J K Rowling and Jim Dale, "Harry Potter and the sorcerer's stone," 2016.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE Signal Processing Society.

[19] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. apr 2015, pp. 5206–5210, IEEE.

[20] Wei Chen and Jack Mostow, "A tale of two tasks: Detecting children's off-task speech in a reading tutor," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 236, pp. 1621–1624, 2011.

[21] Yao Qian, Keelan Evanini, Xinhao Wang, Chong Min Lee, and Matthew Mulholland, "Bidirectional LSTM-RNN for Improving Automated Assessment Of Non-native Children's Speech," in *Proceedings of Interspeech 2017*, Stockholm, 2017, pp. 1417–1421, International Speech Communications Association.

[22] Felix Claus, Hamurabi Gamboa Rosales, Rico Petrick, Horst-udo Hain, and R\"udiger Hoffman, "A Survey about ASR for Children," in *Proceedings of SLaTE*, Grenoble, 2013, pp. 26–30.