INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Using Commercial ASR Solutions to Assess Reading Skills in Children: A Case Report

Timothy Piton[1,2], *Enno Hermann*[3], *Angela Pasqualotto*[1,2], *Marjolaine Cohen*[1,2], *Mathew Magimai.-Doss*[3], *Daphné Bavelier*[1,2]

[1]Université de Genève, Switzerland, [2]Campus Biotech, Switzerland
[3]Idiap Research Institute, Switzerland

timothy.piton@unige.ch, enno.hermann@idiap.ch, angela.pasqualotto@unige.ch, marjolaine.cohen@unige.ch, mathew@idiap.ch, daphne.bavelier@unige.ch

Abstract

Reading is an acquired skill that is essential for integrating and participating in today's society. Yet, becoming literate can be particularly laborious for some children. Identifying reading difficulties early enough is the first, necessary step toward remediation. Here we investigate the opportunities and limitations of integrating commercial, off-the-shelf automatic speech recognition (ASR) services from IBM Watson to ease the administration and evaluation of children's reading assessment tests in French and Italian.

Index Terms: speech recognition, children's speech, reading assessment

1. Introduction

Literacy acquisition is a fundamental milestone in children's education that is reached over years of intensive teaching and practice. The road to reading proficiency can be cumbersome for about 16% of pupils [1]. Facilitating the access to diagnostic tools to help detect or predict reading difficulties is of utmost importance to provide struggling children with personalized guidance and feedback. The proposed work aims to assess whether current commercial automatic speech recognition (ASR) systems can be leveraged to that end.

Diagnostic tools for evaluating reading skills are deeply rooted in classical, paper-and-pencil tasks. In some tasks, children have to read aloud lists of either words, pseudowords (e.g. "bave") or non-words (e.g. "gwoonn"). They have to do so as fast and as accurately as possible, while an experimenter manually records whether they do so correctly or not. Other tasks require the children to identify and manipulate the different sounds of a word in order to produce a new word. For example, the phoneme deletion task instructs children to "Remove the first sound of the word *tomato*." The correct answer being *omato*. In most such tasks the recorded verbal answers are then transcribed phonemically and annotated by an experimenter to extract meaningful information, such as the mistakes made, reading speed, reaction time, and pause time. While still the standard in clinical evaluation of reading, this manual procedure renders the assessment process long and tedious.

Rethinking paper-and-pencil tasks by making use of ASR technology, be it readily available and affordable commercial services or custom-built solutions, has the potential to ease this process and allow children to have an individualized, systematic assessment of their reading-related skills while alleviating the load on educators, clinicians and researchers. Solutions that allow filtering out correct responses, so that clinicians only have to verify likely mistakes would be especially welcome. The application of ASR to monitor children's reading performance [2, 3, 4, 5] or linguistic development [6] is not new, but this field remains nonetheless largely underdeveloped and often limited to the English language. Similar challenges are faced in the field of computer-assisted second language learning for children [7] and adults, where learners could benefit from automated feedback.

Commercial ASR services, such as IBM Watson or Google Cloud ASR, offer relatively reliable transcription results, but are mostly designed to recognize adult speech in a natural context or voice commands. At this time, there are no studies that have shown how such services may be used to assess reading skills in children. A challenge in this goal is not only the need to recognize children's speech, but to do so in the context of list reading or single-item production tasks without semantic context; in fact, the prompts are often pseudowords or non-words that are not present in the ASR training data. In addition, the children are often recorded in noisy classroom or clinical environments.

While custom-built ASR solutions are more flexible, choosing commercial ASR services has a range of advantages. The target users, such as clinicians and educators, often do not have the expertise and resources to develop custom speech technology for their use cases. On the other hand, cloud ASR services offered by big companies are maintained by dedicated experts. Their models are kept up-to-date with recent trends, are trained on larger amounts of data than accessible to most researchers [8], and are built for robustness to handle a variety of down-stream applications.

Some research has been focusing on the development of inhouse ASR services specialized in children's speech recognition with promising results [9, 10], but the resulting solutions are typically limited to specific languages and to corpora of words that do not include items such as non-words and pseudo-words, which are valuable for diagnostic and research purposes.

For these reasons, there is an interest in evaluating the extent to which already existing commercial or custom-built solutions can be used to assess children's speech in the context of reading-related tasks. In this paper, we analyse the performance of the commercial IBM Watson ASR service on Italian and French primary school students taking two reading-related, digitized assessment tasks. For each task, we compare the cloud ASR transcriptions with the standard of the field — manual phonemic transcription by expert clinicians. We also identify the most challenging stimuli in order to uncover the patterns of transcription errors.

2. Methods

In this study we focus on two common reading assessment tasks: word and pseudoword decoding and phoneme deletion tasks. The former requires children to read aloud lists of words and of pseudowords as fast and accurately as possible. The lat-

10.21437/Interspeech.2023-928

ter asks children to listen to a word or pseudoword and repeat it with the first phoneme removed. Before each task, children first complete a practice trial to familiarize themselves with the instructions.

2.1. Recording prompts

Although we are well aware that such materials exist in French and in Italian, new materials had to be generated given our goals (i) to match items features across languages and (ii) to have 3 lists of comparable difficulty so as to allow evaluation at 3 different time points in each language.
Thus, the decoding task comprised three lists of 48 words and 24 pseudowords, and one list of practice items (6 words and 6 pseudowords in French; 12 words and 12 pseudowords in Italian), for a total of 150 words / 78 pseudowords in French and 156 words / 84 pseudowords in Italian. Both words and pseudowords are divided into two categories: *simple* words, which are shorter, have a high frequency of use and a low orthographic complexity, and *complex* words, which are longer, have a lower frequency of use and a higher orthographic complexity. Table 1 shows examples of the stimuli. The practice trials have one list each of simple words and pseudowords, while the main task contains three lists of each.

Table 1: *Examples of items used in the decoding task's test trials.*

Language	Type	Target	Complexity
French	word	nuit	simple
	word	menuisier	complex
	pseudoword	cutice	simple
	pseudoword	pléfantion	complex
Italian	word	pane	simple
	word	inquinamento	complex
	pseudoword	valo	simple
	pseudoword	vusciacope	complex

The initial phoneme deletion task comprised three lists of 18 test items (9 words and 9 pseudowords), and one list of practice items (4 words and 2 pseudowords), for a total of 60 items included in the prompts in each language. Items were selected based on their length, consonant-vowel structure, phonological features of their initial phoneme and frequency of use. They were further equidistributed in length categories, namely short, medium, and long.

Table 2: *Examples of Italian and French items in the initial phoneme deletion task, along with their respective targets and difficulty category.*

Language	Type	Item	Target	Difficulty
French	word	plage	lage	easy
	word	structure	tructure	hard
	pseudoword	razin	azin	easy
	pseudoword	pritunal	ritunal	hard
Italian	word	aspro	spro	easy
	word	sgranato	granato	hard
	pseudoword	detarle	etarle	easy
	pseudoword	sclevosi	clevosi	hard

The vocabulary is representative of clinical reading assessment tasks, being phonetically balanced and of varying complexity to test knowledge of all phonemes in different contexts.

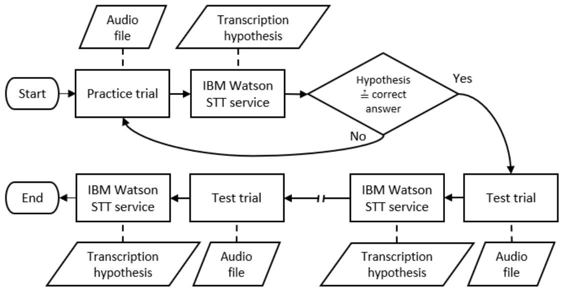

Figure 1: *Flowchart of the decoding task. It is first carried out for words, followed by pseudowords.*

2.2. ASR adaptation and implementation

We selected the IBM Watson cloud ASR[1] solution because it supports adaptation of the language models and adding new words to the vocabulary on paid plans. We use the provided *next-generation multimedia* ASR models for French and Italian, which are optimized for general-purpose audio with a sampling rate of at least 16 kHz. While acoustic model adaptation is not supported, we trained task-specific language models by uploading the list of prompts for each task and language. For the phoneme deletion task, we included both the prompt and the target (i.e., the prompt minus the first phoneme) to allow both to be recognized. We aid the ASR model by passing pronunciation hints for pseudowords through IBM Watson's "sounds like" feature. However, this only allows specifying alternative grapheme sequences rather than phonemic pronunciations, e.g., *piro* for the French pseudoword *pireau*. Therefore, the model might struggle to learn the correct pronunciations for certain pseudowords.

Figure 1 illustrates the trial workflow for the decoding task. During the trial, the Lenovo Tab M10 FHD Android tablet on which the tasks are installed records the participant and saves a WAV audio file for each list of prompts that is sent to the cloud ASR service with the parameters listed in Table 3.

During the practice trials, we compare the transcription hypothesis returned by the cloud ASR service to the correct answer using the word edit distance. If the ASR request fails or if there is a mismatch including deletions or substitutions between the hypothesis and the correct answer, the participant is invited to repeat the practice trial. Insertion errors are not relevant for the assessment and it is in fact common that the recordings contain speech other than the prompted utterances. During the test trials, we store the ASR transcriptions for analysis, but do not provide feedback to the children.

2.3. Experimental setting and procedure

22 French-speaking Swiss (13 females, 7.6 ± 0.5 years old) children and 26 Italian (10 females and 1 unspecified gender, 6.9 ± 0.27 years old) children voluntarily participated in the study. Both the children and their parents have signed a consent form stating they could withdraw from the study at any time, that the collected data would be solely used for scientific purposes and shared only upon their written consent. Ethics approval for the study was required and granted for this study by the Education Research Department of the canton of Geneva.

In the decoding task, participants were successively presented with 5 lists of words (1 practice, 4 tests) and 3 lists of

[1] https://www.ibm.com/cloud/watson-speech-to-text

Table 3: *IBM Watson cloud ASR parameters. The customization weight was set to 1.0 to strongly bias the language models towards the target items. The background audio suppression parameter was increased from the default value of 0.0 to 0.1 in order to silence irrelevant sounds from the environment (e.g. other pupils, etc). Finally, the speech detector sensitivity was increased from 0.5 (default) to 0.6 for better recognition of speech in noisy environments or when the participant speaks softly. We also ask the service to return the start and end timestamps of each utterance and word confidences.*

Parameter	**Value**
Customization weight	1.0
Background audio suppression	0.1
Speech detector sensitivity	0.6
Timestamps	True
Word confidence	True

pseudowords (1 practice, 2 tests), and asked to read them out loud as fast and accurately as possible.

In the phoneme deletion task, children were successively presented with 11 word and 11 pseudoword trials (2 practice, 9 tests) in a pre-defined order and asked to produce out loud the items they heard without their first phoneme.

For both the decoding and the deletion tasks, the child had to press a validation button to end the recording of each trial. All recordings collected during the study remain non-public accordingly to the conditions stated on the consent form.

3. Results

The main aims of this study are to evaluate the extent to which commercial ASR services can faithfully classify the answers provided by the children as correct or incorrect, and to further identify the nature of the children's mistakes. To this end, we compare the manual transcript by clinicians and the ASR transcript with the word edit distance. Any insertions are ignored in order to discard unprompted speech and to allow self-corrections by the children, as clinicians do. We group the remaining tokens into the following categories:

Correctly accepted: Items categorized as correctly pronounced by the clinicians and also present in the ASR transcript.

Correctly rejected: Items categorized as mispronounced by the clinicians for which the ASR output is also different from the target although the exact mistake flagged by each may differ.

Falsely accepted: Mispronounced items as per the clinicians for which the ASR system returns the target.

Falsely rejected: Correctly pronounced items that are not identified in the ASR transcript.

Table 4 shows an example of this evaluation. For each task, we will further indicate which items were most susceptible to be misrecognized by the ASR services. For practical purposes, the precision should be as high as possible, i.e., the number of false accepts should be reduced. This would allow clinicians to safely disregard items identified as correct by the ASR system and manually review only the suggested mistakes.

3.1. Decoding task

Audio files from 21 French- and 25 Italian-speaking children were analysed for the word condition, and 21 French- and 25 Italian-speaking children for the pseudoword condition.

Table 4: *An example prompt with possible manual and ASR transcripts of what the child said. We group the ASR output into insertions (*INS*) that are ignored for our evaluation, correct accepts (*CA*), correct rejects (*CR*), false accepts (*FA*), and false rejects (*FR*).*

Prompt			*brume*		
Manual transcript	br	brume	bain	brune	brume
ASR transcript	br	brume	baume	brume	rume
Result	INS	CA	CR	FA	FR

For words, the participants took the practice trial between 1 and 5 times for French (1.85 ± 1.14), and between 1 and 13 times for Italian (2.8 ± 2.55). The high number of practice trials in some participants is mostly caused by failed requests due to bad Internet connection. For pseudowords, the children took the practice trials between 1 and 10 times for French (2.9 ± 2.69), and between 1 and 7 times for Italian (2.84 ± 1.65).

Figure 2 presents the results of the ASR system on the decoding task. We observe more correct accepts for words than for pseudowords. This is expected because pseudowords are harder both for the children to pronounce and for the ASR system to recognize. Indeed, as they are not present in the acoustic model's training data, the ASR system might not learn the correct grapheme-to-phoneme correspondence from the language model adaptation alone.

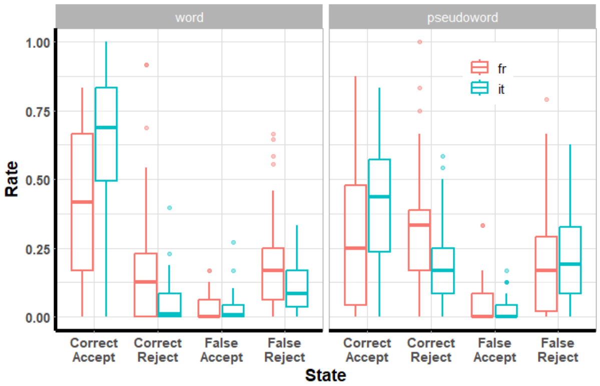

Figure 2: *Decoding Task. Answer rate across participants as a function of our 4 categories of answers. Left panel - word; right panel - pseudoword. French in red and Italian in blue.*

Of special concerns are false accepts. Indeed to help in coding, having clinicians only review the rejected items would gain a significant amount of time, as well as being highly valuable for immediate feedback during the automated practice trials. Table 5 shows the items with the highest false-acceptance rates on the decoding task. Most of the false accepts are due to not identifying minor phoneme differences between prompt and response (bille–bile, compagne–campagne, orchestre–orjestre, prodeglia–prodelia). Additionally, most of these mispronunciations are out-of-vocabulary tokens that the ASR system cannot recognize, so it often returns the target as the closest match.

3.2. Initial Phoneme Deletion task

The audio files of the same 21 French-speaking and all 26 Italian-speaking children were analysed. Figure 3 presents the ASR results on the phoneme deletion task. Unlike for decoding, there is no clear difference between word and pseudoword recognition in this case. This is a direct effect of the

Table 5: *Decoding targets with a false-accept rate (FAR) superior to 0.25, along with the children's responses with the highest occurrence. Items with an asterisk indicate pseudowords.*

Language	Target	FAR	Response
French	bille	0.5	bile(4)
	notice	0.5	notic(3)
	compagne	0.33	campagne(2)
	écaille	0.33	écueille(1), équelle(1)
	oeufs	0.33	oeuf(1), zeuf(1)
	orchestre	0.33	archestre(1), orjestre(1)
	septième	0.33	sepetième(1), sepetiement(1)
	compteur	0.25	competeur(2)
	lesoie*	0.33	lessoie(2)
Italian	arrossamento	0.28	arrostamento(1), arrossa(1)
	prodeglia*	0.25	prodelia(1), prodeia(1)

task, which asks participants to delete the initial phoneme of the prompt. For both words and pseudowords, the resulting recognition target is most likely a token not seen during the initial acoustic model training.

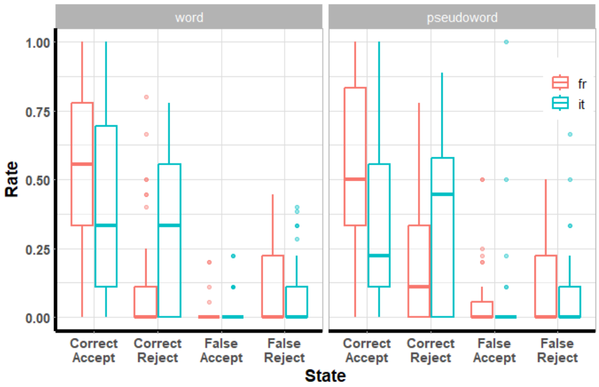

Figure 3: *Phoneme Deletion Task. Answer rate across participants as a function of our 4 categories of answers. Left panel - word; rignt panel - pseudoword. French in red and Italian in blue.*

Again, given our aim to minimize the rate of false accepts, it is of particular interest to identify the items that were recognized as correct by the IBM Watson service when in fact they were not. In Table 6, we present for both languages the words with a false-accept rate of more than 0.25. Again, the most common false accepts are due to not being able to recognize out-of-vocabulary mispronunciations.

4. Discussion

While our results show that cloud ASR services can be leveraged to help the evaluation of children's reading assessment tasks by minimizing the false accept rate, the present work also highlights certain limitations. Pseudowords that were not seen during acoustic or language model training can still be recognized because we adapted the language models with all prompts and target items, although the performance is slightly lower than for words. However, IBM Watson is not able to detect every pronunciation mistake because it is designed to recognize words, not arbitrary phoneme sequences. For example, if a child mispronounces the French word *chambre* /ʃɑ̃bʁ/ as /ʃɑ̃pʁ/, the ASR system could never return *champre* because this token is not in its vocabulary and it might default to *chambre* as the closest match, resulting in a false accept.

Table 6: *Phoneme deletion target items with a false-accept rate (FAR) superior to 0.25, along with the children's responses with the highest occurrence. Items with an asterisk indicate pseudowords.*

Language	Target	FAR	Response
French	aménéon*	0.33	améléon(3)
	olibier*	0.25	olibi(1)
Italian	bronza	0.43	sbronza(1), bronzo(1), pronza(1)
	granato	0.29	gramato(1), ganato(1)

Given our aim to reduce false accepts so that clinicians only have to review rejected items, we propose to combine the cloud ASR model with a dedicated mispronunciation detection system in a two-pass approach. First, the ASR system returns the general word sequence of a spoken utterance. Then, deviations from the target phoneme sequence can be detected with Goodness of Pronunciation [11] or related approaches to flag possible mispronunciations even within the initially accepted set by the ASR system. By reducing the false-accept rate to a negligible level, we should be in a position to allow clinicians to only review Reject classifications from the combined solution. In addition, even for these Reject classifications, our combined ASR solution may already provide initial suggestions for the phonological annotation of errors, again lightening the load on clinicians, as review of proposed mispronunciations is faster than a fully manual phonemic transcription.

5. Conclusions

In this paper we have compared the commercial ASR-based and manual transcripts of French and Italian children's speech utterances acquired during two clinical reading assessment tasks. Our results indicate that while these commercial ASR services do not provide a fine-grained analysis of children's speech by themselves, their transcripts can be used to roughly classify the resulting answers as correct or incorrect, which allows to provide direct feedback during the administration that the overall task is carried out correctly. After language model adaptation, the cloud ASR system also performed well on pseudowords that are not included during the initial ASR training, but are commonly used in reading assessment tasks.

A more detailed analysis of the children's mispronunciations is not possible with an ASR system alone. But in combination with a dedicated mispronunciation detection system, cloud ASR services can alleviate the transcription and classification load on researchers and clinicians. Future work will focus on the implementation and evaluation of this two-pass approach and include trials with more participants.

6. Acknowledgements

This work was partially funded by the NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888 and the IICT flagship, Innosuisse grant agreement No PFFS-21-47 (Enno Hermann).

7. References

[1] S. E. Shaywitz, M. D. Escobar, B. A. Shaywitz, J. M. Fletcher, and R. Makuch, "Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability," *New England Journal of Medicine*, vol. 326, no. 3, pp. 145–150, 1992.

[2] Y. Bai, F. Hubers, C. Cucchiarini, and H. Strik, "ASR-Based Evaluation and Feedback for Individualized Reading Practice," in *Proc. Interspeech*, 2020, pp. 3870–3874.

[3] ——, "An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment," in *Proc. IberSpeech*, Mar. 2021, pp. 11–15.

[4] Y. Bai, F. Hubers, C. Cucchiarini, R. van Hout, and H. Strik, "The Effects of Implicit and Explicit Feedback in an ASR-based Reading Tutor for Dutch First-graders," in *Proc. Interspeech*, 2022, pp. 4476–4480.

[5] K. Reeder, J. Shapiro, J. Wakefield, and R. D'Silva, "Speech recognition software contributes to reading development for young learners of english," *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, vol. 5, no. 3, pp. 60–74, 2015.

[6] S. Dutta, S. A. Tao, J. C. Reyna, R. E. Hacker, D. W. Irvin, J. F. Buzhardt, and J. H. Hansen, "Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments," in *Proc. Interspeech*, 2022, pp. 4322–4326.

[7] L. Ling and W. Chen, "Integrating an ASR-based translator into individualized L2 vocabulary learning for young children," *Education and Information Technologies*, 2022.

[8] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wu, "BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, Oct. 2022.

[9] A. C. Kelly, E. Karamichali, A. Saeb, K. Vesely, N. Parslow, G. M. Gomez, A. Deng, A. Letondor, N. Mullally, A. Hempel, R. O'Regan, and Q. Zhou, "SoapBox Labs Fluency Assessment Platform for Child Speech," in *Proc. Interspeech*, 2020, pp. 488–489.

[10] A. C. Kelly, E. Karamichali, A. Saeb, K. Vesely, N. Parslow, A. Deng, A. Letondor, R. O'Regan, and Q. Zhou, "Soapbox Labs Verification Platform for Child Speech," in *Proc. Interspeech*, 2020, pp. 486–487.

[11] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, Feb. 2000.