Human Interaction and Emerging Technologies (IHIET 2023), Vol. 111, 2023, 213–223
https://doi.org/10.54941/ahfe1004027

Digital Primer Implementation of Human-Machine Peer Learning for Reading Acquisition: Introducing Curriculum 2

Daniel Devatman Hromada[1,2] and Hyungjoong Kim[2]

[1]Institute for Time-Based Media, Faculty of Design, Berlin University of the Arts, Germany
[2]Einstein Center Digital Future, Berlin, Germany

ABSTRACT

The aim of the digital primer project is cognitive enrichment and fostering of acquisition of basic literacy and numeracy of 5 – 10 year old children. Here, we focus on Primer's ability to accurately process child speech which is fundamental to the acquisition of reading component of the Primer. We first note that automatic speech recognition (ASR) and speech-to-text of child speech is a challenging task even for large-scale, cloud-based ASR systems. Given that the Primer is an embedded AI artefact which aims to perform all computations on edge devices like RaspberryPi or Nvidia Jetson, the task is even more challenging and special tricks and hacks need to be implemented to execute all necessary inferences in quasi-real-time. One such trick explored in this article is transformation of a generic ASR problem into much more constrained multi-class classification problem by means of task-specific language models / scorers. Another one relates to adoption of "human machine peer learning" (HMPL) strategy whereby the DeepSpeech model behind the ASR system is supposed to gradually adapt its parameters to particular characteristics of the child using it. In this article, we describe execution of first exercise by means of which the Primer assisted author's 5-year-old daughter in increase of her syllable-reading competence. The pupil went through sequence of exercises composed of evaluation and learning tasks. Consistently with previous HMPL study, we observe increase of both child's reading skill as well as of machine's ability to accurately process child's speech.

Keywords: Personal primer, Educational instrument, Digital reading acquisition assistant, Speech recognition, Edge computing, Human-machine peer learning, Task-specific scorers

INTRODUCTION

This article provides insight into first results issued out from a Personal Primer (PP) project inspired by Stephenson's (2003) Young Lady's Illustrated Primer proposal and initiated in 2018 by a publication of a roadmap leading to a post-smartphone, book-like, do-it-Yourself *Bildunginstrument* (Hromada, 2019b). Ideally, the PP aims to attain both education-about-digital as well as education-with-digital objectives (Hromada, 2020). These are:

- education-about-digital objective: increase digital competences of older students so that they are able to repair or ameliorate existing Primers or construct their new copies
- education-with-digital objective: develop an open-source software suite and open educational resource (OER) corpus which could help elementary school pupils successfully enter the world of basic literacy

While non-negligible amount of work has already been executed in making the education-about-digital objective real and concrete (Hromada et al., 2020) by means of combination of well-known off-the-shelf components (e.g. RaspberryPi Zero, e-ink display, c.f. Figure 1), this article focuses on research & development practices by means of which we aspire to attain the education-with-digital objective.

That is, on practices which whose proper implementation in the Digital Primer (DP) branch of the PP project[1] may help 5–10 year old children learn how to write and – this is the main topic of this article – *learn how to read*.

Figure 1: Left-hand and right-hand version 1 prototypes of Personal Primer (PP) artifact with integrated speech capabilities, e-ink displays, circadian module (Hromada, 2021), solar panel and touchless gesture command & control interface.

Digitally Supported Reading Acquisition

Reading is a culturally determined skill whose acquisition can be fostered or inhibited by learner's exposure to appropriate respectively inappropriate social, pedagogic and instrumental environment. Reading is also a fundamental skill in a sense that it is a *condition sine qua non* for acquisition of other skills and is considered to be "*a central prerequisite for a successful educational biography and social participation*" (Alscher et al., 2022).

[1]We use the term Personal Primer (PP) to denote the physical, book-like hardware artefact and the term Digital Primer (DP) to denote the software used by both PP as well as by diverse web projects and apps like fibel.digital, palope.digital etc. Should we use the simple term "Primer", we refer to both PP and DP in the same time.

Unfortunately, in Germany - where the PP has first took shape not only as a science-fiction proposal but also as a tangible project - reading competences of pupils seem to steadily decline since 2011. What's worse, recent school panel study indicates additional "*substantial decline in reading achievement*" caused by CoVid-related lockdown policies (Patrinos et al., 2022). Thus, measures to counteract the trend seem to be of non-negligible societal importance.

It is generally believed that one such counter-measure may be obtained by implementation and deployment of digital learning tools – or digital reading acquisition assistants (DRAAs)- which foster pupil's acquisition of reading skill there, where other tutor – ideally a human teacher, parent or peer – is not available or unable to help the child to master a cognitively challenging task of "learning how to read".

For example, promoters of Microsoft's Schlaumäuse claim that use of their app may facilitate reading acquisition: as a result the app is currently used in more than 6000 Kindergartens in Germany. However, independent empiric evidence for such claims remains sporadic and criticism of an app based on sorting exercises filled with colored, animated, and high-pitched sounds seems to be appropriate both on cognitive-, didactic- as well as media-pedagogic grounds.

Situation is not very much different in case of Google's "Read Along" app where an internal "speech recognition" mechanism should give an avatar named Diya ability to recognize student's reading difficulties. Unfortunately for the pupil, Diya's speech-processing capabilities are still not very much accurate. This results in non-negligible amount of false negatives – where wrong lecture is recognized as correct – and even worse, false positives – where correct lecture is labeled as erroneous (Seidler and Hromada, 2021).

Inspite of these and other privacy-related drawbacks, the Read Along DRAA is still considered as efficient means of increasing student's reading achievements by those studies which more closely analysed the efficiency of the "Read Along" software (Yoon, 2022).

Reading Acquisition and Automatic Speech Recognition

Reading is, *in essentiam*, a process of translation of graphemic, textual sequences into their phonetic representations. In younger learners, this process is fully externalized and has a form of "reading out loud"; later, with higher level of mastery, process is more internalized and the phonetic representation gets actualized only virtually, in form of an "inner speech" (Jones, 2009) soundlessly occuring within reader's mind.

Speech and spoken word thus play a fundamental role in reading acquisition – both in learning and practice as well as evaluation of reading competence, diagnosis of potential reading difficulties or even their therapy (Davis and Braun, 2011). For this reason, a DRAA endowed with an accurate automatic speech recognition (ASR) component is, *ex vi termini*, bound to be more natural and efficient a tool than DRAA whose ASR component is inaccurate or altogether absent.

While highly accurate ASR systems exist for many languages (Ravanelli et al., 2021) they are still strongly biased towards accurate processing of healthy adult voices. However, in reading acquisition or reading fostering scenarios one deals with subjects whose utterances of sequences-to-be-read exhibit peculiar characteristics. Repetitions, pauses, ligatures, interjections, stuttering-like (Bayerl et al., 2023) artefacts and other phenomena accompany the process of reading acquisition. Therefore, it is fairly certain that in order to train an ASR model useful for DRAA, one first needs to collect and create voice recording datasets of such phenomena. In datasets on which all current state-of-the-art (SOTA) publicly available ASR systems were trained, such phenomena are often rare or even completely absent.

Child Speech Recognition

Another reason why ASR systems are – as of 2023 – not yet accurately implemented in publicly available DRAAs is caused by the trivial observation that majority of subjects who learn how to read are children.

There is higher variability among children voices than among adult voices. This is due to combination of multiple factors such as differences in size and anatomy of vocal tract – including ontogenetic phenomena like teeth eruption and teeth change; phonological competences; level of motric control etc (Beckman et al., 2017).

As a result, even the most basic features of a child's voices are different from those of an adult– fundamental frequencie[2] are different and so are Mel-frequency cepstral coefficient (Levitan et al., 2016) features. That is, features which provide very input vectors for majority of dominant ASR models.

Or, as (Rumberg et al., 2022) state: "*datasets for children's speech that are publicly available are very scarce, especially when specific languages and/or connected natural speech are targeted*".

Thus, in Federal Republic of Germany (Rumberg et al., 2022), as well as in few other countries like Italy (Gretter et al., 2020), Australia (Ahmed et al., 2021) or linguistic communities like Kannada (Ramteke et al., 2019) certain initiatives have been undertaken to create datasets of children speech and, ideally, make them available to wider research community.

Nonetheless, in spite of growing amount of resources, current SOTA systems for child speech recognition are still far from being perfectly accurate. For example, the authors of the kidsTALC repository of typically developing monolingual German children report 26.2 % word-error-rate (WER) when evaluating the testing section of their dataset with SOTA neural-network architectures includes in the speech processing framework SpeechBrain (Ravanelli et al., 2021).

Thus, additional data needs to be collected or other novel means of problem complexity decrease or accuracy increase need to be found, implemented and tested. Two such means labeled as "human-machine peer learning" and "task-specific language models" ar described in the next section.

[2]Fundamental frequency of male voice is 112.0 Hz, fundamental frequency of female voice is 195.8 Hz (Oliveira et al., 2021), fundamental frequency of boy voice is 250.0 Hz and fundamental frequency of girl's voice is 244.0 Hz (Linders et al., 1995).

DIGITAL PRIMER

Digital Primer (DP) implements two such innovations: "human-machine peer learning" (HMPL) and "task-specific language models". While the first one aims for increase of accuracy by gradual adaptation of the ASR model to a specific human learner, the second one is based on the idea that majority of reading acquisition exercises can be conceptualized as domain-constrained sub-problems of a much more challgenging generic ASR problem.

DP is a web-based, progressive web app (PWA) complement to the physical artefact. Accessible at the URL fibel.digital as a frontend to a complex *knowledge graph at its backend*, some of its features and contents – e.g. "memory game" ; player for multi-font, multi-voice, multi-accent collections of audiotextual OERs; active folios etc. (see Figure 2) - can be used without authentication by any visitor of the website.

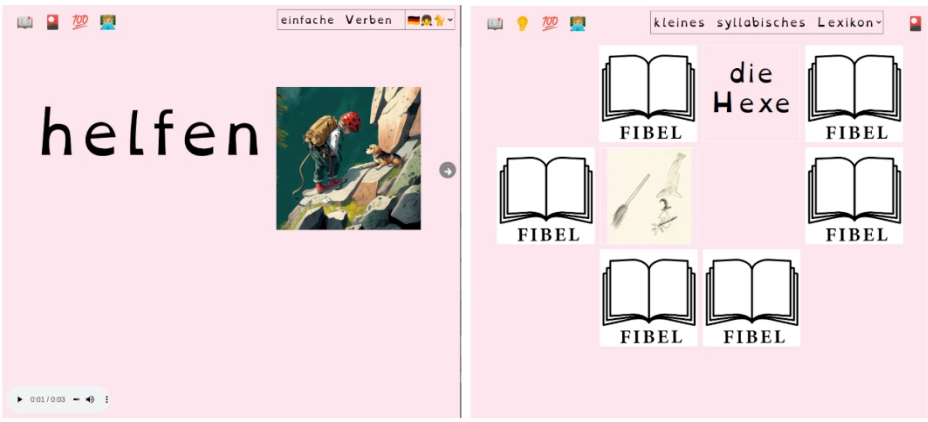

Figure 2: Multi-voice audiotext player (left); and memory game (right) features of DP's progressive web app are available even to non-authenticated users.

Human-Machine Peer Learning

Primer is a first system which implements the concept of "human-machine peer learning" for the purpose of reading acquisition. Introduced in (Hromada, 2022) with the simple statement "*humans and machines can learn from each other*", HMPL provides a paradigm for construction of such human-machine learning curricula from which both humans as well as machines benefit.

In a first empiric HMPL "Curriculum 1" (HMPL-C_1) study which focused on extending foreign language vocabulary for humans while simultaneously increasing speech-recognition accuracy of "artificials", we "*observed increase in amount of matches between expected and predicted labels caused both by increase of human learner's vocabulary, as well as by increase of recognition accuracy of machine's speech-to-text mode*" (Hromada and Kim, 2023).

In the scope of the Primer project, the primary focus is not on acquisition of a new language but rather on acquisition of a new semiotic system like

Table 1. HMPL-table for reading acquisition (Curriculum 2) for human learner individual I and artificial utterence-processing tutor U.

Curriculum 2	**I**	**U**
Role	Human learner *I*	Machine *U*
Curricular objective	ability to read script S	accurate transcription of *I*'s speech
Exercise 1		
Skill	Π = reading C+V syllables	σ = accurate transcription of syllabic sequences uttered by *I*
Prior knowledge	sound	sound files with corresponding labels
Input	text	speech
Output	speech	text

script S. For this reason, a second HMPL Curriculum (HMPL-C_2) is hereby proposed.

HMPL-C_2 is a sequence of exercises at the outset of which the human individual I acquires the ability to read texts written in script S while the artificial utterance-processing tutor U [3] acquires the ability to process I's speech evermore accurately. This articles focuses on initial stages of exercise 1 (E_1) whereby I should learn how to read simple syllables.

Task-Specific Scorers

Modern ASR systems often include combination of ideas known as "connectionist temporal classification" (CTC) and "beam search" (Graves, 2012). This allows to extend a so-called "acoustic model" – which, for every time frame T included in the speech signal yields a distribution of most probable phonemes – with a "language model" which encodes information about probabilities of sequences of the target language. Stated in other terms, the "language model" – sometimes also called "scorer" – guides the process where sequences of candidate phonemes concatenate into sequences of candidate words and phrases.

In a generic ASR context, such a language model is a complex data structure encoding transition probabilities among dozens of thousands words of the target language. However, in cases where target language is expected to be more constrained, usage of such models is often unnecessary or even counter-productive. And such is, indeed, also the case where one deals with learners – like small children, for example, whose language is a simplified protovariant of its future adult state. For example, having a language model which includes the word "ridge" and use it in an ASR system for speakers whose lexicon contains word "bridge" but not "ridge" may lead to decrease

[3] Contrary to our previous publications where we used letters H and M to denote human-, resp. machine-learners, we the letters "I" resp. "U" (Buber, 1923) to denote human- resp. machine- peers of the HMPL process.

of accuracy of such system, in spite of the fact that the scorer is bigger and one would thus naively expect it to be better.

In majority of reading exercises which are – or will be – included in the Primer, one already knows the text which is to be read and thus, one already knows what utterances could be considered as correct lectures and which not. Thus, the task is much more constrained and generic ASR / STT system would be an overkill. For this reason, we use for every specific exercise - like vowel or syllable recognition – a specific scorer which is constraining the beam search to restricted amount of exercise-relevant answers. In other terms, exercise-specific scorers allow us to easily transform a generic acoustic model into multi-class classifier. By significantly constraining the search space of plausible solutions [4], the accuracy of such domain-constrained ASR system is expected to increase.

PRELIMINARY STUDY

This section describes preliminary results obtained after three HMPL iterations between one particular child and the DP prototype.

Learner 1

Human subject involved in the study – and labeled hereby as Learner 1 (L_1) - is a 5-year old – pre-school - daughter of the main author of this article. Apart to read and write its own name, name of its sister and words "mama" and "tato", very limited alphabetic competence has been attested previous to this study.

Being aware that performing studies with an own child can be a source of significant bias, the author is nonetheless convinced that it is an only ethically viable approach in cases of highly experimental AIED research[5] where ethics cannot, should not and must not be ignored. Thus, with a current study, we join such developmental psychologues such as Piaget and psycholinguists such as Tomasello (Tomasello, 2005) whose research also involved observations of cognitive development of their own children.

Sessions

In the course of the preliminary study, we have executed three HMPL-C_2 sessions on days 1, 3 and 5. Each session consisted of human-testing phase followed by a mutual human-machine learning phase. During the human-testing phase, L_1 was asked to read out loud syllabic sequences displayed by the DP interface. Evaluated sequences consisted of 5 repetitions of syllables started with occlusive labial consonant M or B and followed by the vowel A, E, I, O or U, thus yielding sequences from "MA MA MA MA MA" to "BU BU BU BU BU".

[4] C.f. pages 68 and 152 of (Hromada, 2019a) further references supporting the "less is more" hypothesis in both developmental and computational linguistics.

[5] C.f. the recent pre-print (Hromada, 2022) or its future derivatives for closer discussion of ethical aspects of testing, deploying and evaluation of child-directed artificial intelligence in education (AIED) technologies.

During the learning phase, L_1 first listened to the sequence which was also displayed as the text on the screen, and then was asked to repeat what she just heard. In order to strengthen both L_1 's as well as model's ability to correctly process most important building blocks of language – vowels -, the ten CV-sequences used in the human-testing phase were, during the learning phase, also accompanied by sequences of five repeated vowels A, E, IE, O, and U.

During the learning phase, itt was also pointed out to the child that the sounds she just heard correspond to letters which she sees and that when she repeats the recent utterance, she actually "reads".

Speech recordings collected during the learning phase subsequently provided input for the acoustic-model fine-tuning process. Speech recordings collected during the human-testing phase weren't included in the machine learning.

Models

All models evaluated in this article are based on Mozilla's implementation of DeepSpeech (DS) architecture (Agarwal and Zesch, 2019). The baseline model is publicly available DS model which was trained on Common Voice Data. We subsequently fine-tuned the baseline model with data provided by kidsTALC project (Rumberg et al., 2022), thus obtaining the model which we hereby label as KI^6ds^0.

Subsequently, after termination of every mutual-learning phase, the $KIds^0$ model has been incrementally fine-tuned with data just provided by L_1. One single epoch of training took place, with 0.0001 learning rate. In such a manner, we obtained three additional models: $KIds^{L1\text{-}1}$ is $KIds^0$ fine-tuned on data collected during the human-machine learning phase on day 1; $KIds^{L1\text{-}3}$ is $KIds^{L1\text{-}1}$ fine-tuned with data collected during the human-machine learning phase on day 3 and $KIds^{L1\text{-}5}$ is $KIds^{L1\text{-}3}$ trained with data collected during the human-machine learning phase on day 5.

Evaluation Metrics

We evaluated results of the preliminary study by means of a standard "word error rate" (WER) metrics. Sequences of five vowels resp. CV syllables which were displayed by DP were considered to provide the "reference"; output of the model yielded the hypotheses. Full match between reference and the hypothesis[7] yields $WER_{ideal}=0$, zero correspondence yields $WER_{worst}=1$. If only one among five reference syllables matched, WER = 0.2, if two matched WER = 0.4 etc.

RESULTS

Results of the preliminary 3-session study focusing solely on L_1 are summarized in Table 2. Decrease of mean WER between columns points to potential increase in L_1's syllable-reading competence. Decrease of WER between rows indicates that accuracy of the ASR model increased as well – this may indicate

[6]Note that "KI" is German equivalent to "AI".

[7]In HMPL, error can be caused not only by artificial but also by human factor.

Table 2. Mean word error rates yielded by five different ASR models (rows) & three human-testing (columns) sessions which took place on days 1, 3 and 5 of HMPL-C2-E1.

	Day 1	Day 3	Day 5
DeepSpeech_DE	0.96	0.84	0.64
KIds^{0}	0.74	0.72	0.68
KIds$^{L1\text{-}1}$	0.69	0.78	0.44
KIds$^{L1\text{-}3}$	0.69	0.8	0.52
KIds$^{L1\text{-}5}$	0.69	0.74	0.48

that model Kids^{L1} initiated gradual adaptation of its parameters to peculiar properties of L_1's voice and her verbal behaviour.

CONCLUSION

Creating an artificial assistant which helps *the* child learn how to read and write is not an easy task. Learning how to read is a laborious task which necessitates patience and attention. And *the* child cannot and should be treated neither as a "user", nor as a "consumer": it is very nature of being *the* child to be wild and vivid; to resist to assimilate dry adult-made semiotic systems. And it is good so.

Additionally, phonic properties of children voices are different from those of adults and children voice datasets are limited. Thus, it is not surprise that as of 2023, there exists no publicly available ASR model which could accurately and reliably process child speech.

In this article, we have introduced two hacks & tricks how the problem can be partially bypassed. One relates to transformation of a generic ASR problem into a sort of extended multi-class classification problem by means of extending a generic acoustic model with a domain-specific, minimalist language model ("scorer"). Another relates to the method of human-machine peer learning (HMPL) whereby the artificial utterance-processing tutor **U** incrementally and gradually adapts its parameters to a particular learner, a human individual **I**. In concrete terms, we have shown that after three sessions focusing on acquisition of grapheme-vowel and CV-bigrapheme correspondences had lead, in case of one particular **I – U** couple, to decrease of WER from 96% to 48%.

Still, given the fact that the individual **I** = L_1 = a daughter of main author of this article, biases cannot be excluded and any generalization concerning the usefulness of HMPL-based assistance to reading acquisition - as currently implemented in the DP project - has to be postponed to further studies. For it is only such studies – where learning shall be executed not only by subjects healthy and privileged, but also by subjects handicapped, discriminated and traumatized - which could show whether the Primer project hereby introduced is yet-another-corporate-snake-oil or a do-it-Yourself Bildunginstument and an embedded AI project containing all necessary ingredients to shine with its own inner light.

ACKNOWLEDGMENT

Acknowledgment goes to prof. Christa Röber and didactic community around the syllable-based primer concept "Zirkus Palope" for their kind, patient and obstacle-transcending work on alphabetisation of children from all social layers and migration backgrounds.

REFERENCES

Agarwal, A., Zesch, T., 2019. German End-to-end Speech Recognition based on DeepSpeech. In *KONVENS*.

Ahmed, B., Ballard, K., Burnham, D., Sirojan, T., Mehmood, H., Estival, D., Baker, E., Cox, F., Arciuli, J., Benders, T. and Demuth, K., 2021. AusKidTalk: an auditory-visual corpus of 3-to 12-year-old Australian children's speech. In Annual Conference of the International Speech Communication Association (22nd: 2021) (pp. 3680–3684). International Speech Communication Association.

Alscher, P., Ludewig, U. and McElvany, N., 2022. Civic education, teaching quality and students' willingness to participate in political and civic life: Political interest and knowledge as mediators. Journal of youth and adolescence, 51(10), pp. 1886–1900.

Bayerl, S. P., Gerczuk, M., Batliner, A., Bergler, C., Amiriparian, S., Schuller, B., Nöth, E. and Riedhammer, K., 2023. Classification of stuttering–The ComParE challenge and beyond. Computer Speech & Language, p. 101519.

Buber, M., 1923, Ich und Du.

Beckman, M. E., Plummer, A. R., Munson, B. and Reidy, P. F., 2017. Methods for eliciting, annotating, and analyzing databases for child speech development. Computer speech & language, 45, pp. 278–299.

Davis, R. D. and Braun, E. M., 2011. The gift of dyslexia: why some of the brightest people can't read and how they can learn. Souvenir Press.

Desplanques, B., Thienpondt, J. and Demuynck, K., 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Graves, A., 2012. Connectionist temporal classification. Supervised sequence labelling with recurrent neural networks, pp. 61–93.

Gretter, R., Matassoni, M., Bannò, S. and Falavigna, D., 2020. TLT-school: a corpus of non native children speech. arXiv preprint arXiv:2001.08051.

Hromada, D. D., 2019a. Prolegomena Paedagogica: Intramental Evolution and Acquisition of Toddlerese. Union, Berlin.

Hromada, D. D., 2019b. After smartphone: Towards a new digital education artefact. Enfance, (3), pp. 345–356. doi: 10.3917/enf2.193.0345.

Hromada, D. D., 2020. Digital education:" education-with"/" education-about" distinction and the teleological definition. On digital education, pp. 1–10. doi: 10.25624/kuenste-1326.

Hromada, D. D., Seidler, P. and Kapanadze, N., 2020. Bauanleitung einer digitalen Fibel von und für ihre Schüler. Mobil mit Informatik, 9, p. 37.

Hromada, D. D., 2021. Three Principles, 2 Sub-principles and One Magic Wand for Harm Minimization and Prevention of Technological Addiction in Human Children. Educational Innovations and Emerging Technologies, (1), pp. 48–57. doi: 10.35745/eiet2021v01.01.0005.

Hromada, D. D., 2022, July. Foreword to Machine Didactics: On Peer Learning of Artificial and Human Pupils. In Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks,

Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part II (pp.387–390). Cham: Springer International Publishing.

Hromada, D. D. and Kim, H., 2023. Proof-of-Concept of Feasibility of Human-Machine Peer Learning for German Noun Vocabulary Acquisition. Frontiers in Education. doi: 10.3389/feduc.2023.1063337

Jones, P. E., 2009. From 'external speech'to 'inner speech'in Vygotsky: A critical appraisal and fresh perspectives. Language & Communication, 29(2), pp. 166–181.

Levitan, S. I., Mishra, T. and Bangalore, S., 2016, May. Automatic identification of gender from speech. In Proceeding of speech prosody (pp.84–88). Semantic Scholar.

Linders, B., Massa, G. G., Boersma, B. and Dejonckere, P. H., 1995. Fundamental voice frequency and jitter in girls and boys measured with electroglottography: influence of age and height. International journal of pediatric otorhinolaryngology, 33(1), pp. 61–65.

Oliveira, R. C., Gama, A. C. and Magalhães, M. D., 2021. Fundamental voice frequency: acoustic, electroglottographic, and accelerometer measurement in individuals with and without vocal alteration. Journal of Voice, 35(2), pp. 174–180.

Patrinos, H. A., Vegas, E. and Carter-Rau, R., 2022. An analysis of COVID-19 learning loss.

Ramteke, P. B., Supanekar, S., Hegde, P., Nelson, H., Aithal, V. and Koolagudi, S. G., 2019. NITK Kids' speech corpus. emotion, 491, pp. 4–15.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J. and Chou, J. C., 2021. SpeechBrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624.

Rumberg, L., Gebauer, C., Ehlert, H., Wallbaum, M., Bornholt, L., Ostermann, J. and Lüdtke, U., 2022. kidsTALC: A Corpus of 3-to 11-year-old German Children's Connected Natural Speech. In Proceedings INTERSPEECH.

Seidler, P. and Hromada, D., 2021, Lesekompetenz und künstliche Intelligenz, Deutsche Kinderhilfe e. V., pp. 45.

Stephenson, N., 2003. The diamond age: Or, a young lady's illustrated primer. Spectra.

Tomasello, M., 2005. Constructing a language: A usage-based theory of language acquisition. Harvard university press.

Yoon, T., 2022, A Study on the Effects of Read Along by Google with Primary ELLs' Pronunciation and Affective Domains, Jour. of KoCon.a, 22(10), pp. 437–444.