sciendo
DOI 10.2478/jazcas-2025-0020

DEVELOPMENT OF A DATABASE AND MODELS FOR CHILDREN'S SPEECH IN THE SLOVAK LANGUAGE FOR SPEECH-ORIENTED APPLICATIONS

JÁN STAŠ[1] – STANISLAV ONDÁŠ[2] – MATÚŠ PLEVA[3] –
MATEJ HORVÁTH[4] – RICHARD ŠEVC[5] – PATRIK MICHALANSKÝ[6]

[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0001-7403-0235)

[2]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-0075-3788)

[3]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-4380-0801)

[4]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia

[5]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia

[6]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia

STAŠ, Ján – ONDÁŠ, Stanislava – PLEVA, Matúš – HORVÁTH, Matej – ŠEVC, Richard – MICHALANSKÝ, Patrik: Development of a Database and Models for Children's Speech in the Slovak Language for Speech-oriented Applications. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 223 – 233.

Abstract: Children's speech differs significantly from adult speech due to physiological and cognitive developmental factors. Key differences include higher pitch, a shorter vocal tract, greater formant frequencies, slower speaking rates, and greater variability in pronunciation and articulation. These differences result in acoustic mismatches between children's and adult speech, making traditional automatic speech recognition models trained on adult speech less effective for children. Additionally, linguistic differences, such as limited vocabulary and evolving grammar, further contribute to this challenge. This paper focuses on the creation of a children's speech database for the low-resource Slovak language. This database has been used to train acoustic models for the automatic recognition of spontaneous children's speech in Slovak. In this research, we compared three different approaches to speech recognition, with self-supervised learning achieving results comparable to similar studies in this area, despite using relatively small amounts of training data.

Keywords: acoustic model, automatic speech recognition, data augmentation, children's speech, speech database

1 INTRODUCTION

Automatic speech recognition (ASR) for children remains a challenging area due to fundamental differences in various acoustic and linguistic aspects between children's and adults' speech. Acoustically, children's speech is characterized by a higher fundamental frequency (F0), increased formant frequencies (F1–F3), slower speaking rates, and greater variability in articulation due to their developing vocal tracts and speech motor control (Patel 2014; Shivakumar 2020; Lu 2022). Linguistically, children's speech exhibits greater variability in pronunciation, increased disfluencies, and evolving phonetic structures as they develop (Yeung 2018). Word pronunciation can be inconsistent due to incomplete language acquisition. Additionally, children's shorter vocal tracts result in different resonance characteristics, making the direct adaptation of adult-trained ASR models ineffective (Gerosa 2009; Lu 2022). These factors collectively lead to higher Word Error Rates when adult-trained ASR systems are applied to children's speech (Sobti 2024).

To address these challenges, various methods have been explored in recent years. Traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) approaches have been widely used, but recent advances favor Deep Neural Networks (DNNs), End-to-End architectures, transformer-based or self-supervised learning models such as Wav2Vec 2.0, which have significantly improved children's ASR performance (Bhardwaj 2012; Shivakumar 2020; Lu 2022; Sobti 2024). Adaptation techniques such as Vocal Tract Length Normalization, Feature-space Maximum Likelihood Linear Regression, Subglottal Resonance Normalization, and Speaker Adaptive Training have been applied to reduce acoustic mismatches between adult and child speech (Patel 2014; Yeung 2018). Transfer learning from adult ASR models to child-specific models has also proven effective, particularly when adapting both acoustic and pronunciation models (Shivakumar 2020). Additionally, data augmentation approaches such as Speed Perturbation and Spectral Augmentation have been used to simulate children's speech. Self-supervised learning has also emerged as a promising approach to mitigate the scarcity of children's speech data, with fine-tuning pre-trained adult models on a small amount of children's speech yielding significant improvements (Lu 2022; Sobti 2024).

Despite these advancements, research gaps remain, including persistently high WER, limited robustness to spontaneous speech, and poor performance in low-resource languages. A significant challenge is the lack of large, publicly available multilingual child speech corpora, which hinders model training and limits the generalizability of ASR systems (Yeung 2018; Sobti 2024). Furthermore, most studies focus on read speech, whereas spontaneous speech recognition remains underexplored (Gerosa 2009). Moreover, speech variability across different child age groups suggests that a single ASR model may not be effective for all children

(Yeung 2018). Age-specific models and improved data collection methods are crucial to enhancing ASR systems' accuracy and adaptability.

Since existing adult-trained ASR models struggle with the unique acoustic and linguistic characteristics of children's speech, building new child-specific datasets and developing tailored models are essential steps toward bridging the performance gap in children's speech recognition. For these reasons, we decided to create a children's speech database for the Slovak language to expand the existing range of languages. With the help of this data, we aim to train acoustic models (AMs) applicable to recognizing children's spontaneous speech, as well as to the design and development of other speech-oriented applications, such as children's speech synthesis and human-machine or human-robot interfaces.

2 DATABASE OF CHILDREN'S SPEECH

2.1 Existing databases of children's speech

Children's speech databases are essential for advancing ASR research. This section provides an overview of key corpora, highlighting their design and data collection methods.

The CMU Kids Corpus is a database of children's read-aloud speech recorded in American English. It includes speech from 76 children aged 6 to 11, with 24 male and 52 female speakers. The corpus consists of 5,180 utterances created to train ASR models for the LISTEN project at Carnegie Mellon University. The dataset is divided into two subsets: SUM95, which contains speech from proficient readers recorded in summer camps, and FP, which includes speech from children at risk of developing poor reading skills (Eskenazi 1997).

The OGI Kids' Speech Corpus is a database of children's speech in American English, consisting of recordings from approximately 1,100 children, ranging from kindergarten to grade 10. The corpus includes both prompted and spontaneous speech, with a balanced number of male and female speakers across different grade levels. Data collection involved children reading words and sentences displayed on a screen while synchronized with an animated character (Baldi), whereas spontaneous speech was elicited through conversational prompts (Shobaki 2000).

The PF-STAR Children's Speech Corpus is a multilingual database containing speech recordings from 611 children in British English, German, Italian, and Swedish. It includes both native and non-native speech, featuring read, imitated, spontaneous, and emotional speech recordings, covering an age range from 4 to 14 years. Data collection employed various methodologies, including scripted reading tasks and the AIBO method, in which children interacted with a robot to elicit natural and emotional speech (2005).

The Child Language Data Exchange System (CHILDES) is a database designed for studying child language acquisition. It includes recordings and transcripts in over

20 languages, making it a crucial resource for researchers. The corpus contains data from thousands of children and caregivers, though the distribution of male and female speakers varies across individual subcorpora. It was built by collecting, transcribing, and annotating naturalistic parent-child interactions, recorded in home environments (Sanchez 2019).

The My Science Tutor (MyST) corpus is one of the largest collections of children's conversational speech, containing 393 hours of speech data from 1,371 students in grades 3–5. The corpus is in English and consists of 228,874 utterances recorded during 10,496 virtual tutoring sessions, with equal participation from male and female students. The data was collected through structured spoken dialogues between students and a virtual science tutor, in which students answered science-related questions in a strict turn-taking system (Pradhan 2024).

The study by Claus et al. (Claus 2013) provides a comprehensive survey of children's speech databases, which are essential for ASR research. It identifies and describes a total of 34 databases, primarily in English, with some available in German, Italian, Swedish, and other languages – unfortunately, excluding Slovak. Most databases contain read or spontaneous speech from children aged 6 to 18 years, with significantly fewer resources for younger children. Preschool speech databases are particularly scarce due to recording challenges, as young children cannot read and have short attention spans. The study highlights the need for more extensive and higher-quality speech databases, especially for children under the age of six.

2.2 Building of the Slovak children's speech database

We began working on the creation of the speech database as early as 2018. A portion of the database, consisting solely of excerpts from children's speech taken from the eight parts of the series Dads ('*Oteckovia*'), was published in (Pleva 2019). '*Oteckovia*' was a Slovak family daytime series broadcast on TV Markíza, depicting the lives of four young men—fathers—each struggling with the role of parenthood in his own way. It is an adaptation of the 2014 Argentine telenovela '*Señores Papis*'. Although the speech in the series has a spontaneous character, it is still a spoken script which we do not consider fully authentic.

For this reason, we proceeded to transcribe more authentic speech from child speakers. We focused on two programs: the TV show '*Táraninky*', broadcast by Slovak television RTVS between 2020 and 2023, and the radio show '*Rozhlasové leporelo*', aired on Rádio Regina between 2021 and 2023. Both programs feature speech interactions in which adult speakers ask various types of questions, and child respondents provide answers. '*Táraninky*' was a children's talk show covering various topics, hosted by Marián Čekovský and his little guests. '*Rozhlasové leporelo*' was a children's radio show focused on developing thinking and creativity, encouraging spontaneous reactions and communication skills.

2.3 Data preprocessing and transcription

As already mentioned, the process of acquiring, pre-processing and transcription of the subcorpus of the series ‘*Oteckovia*’ is described in more detail in (Pleva 2019).

Audio recordings of the shows ‘*Táraninky*’ and ‘*Rozhlasové leporelo*’ were obtained from the RTVS online archive[1] and processed in similar way. A total of 70 episodes of the radio show ‘*Rozhlasové leporelo*’ were analyzed, with 36 episodes discarded due to excessive background noise.

Next, only speech segments containing child speech were isolated using the Audacity tool[2], while segments featuring older children, adults, or other irrelevant audio content were removed. A three-second pause was inserted between individual segments of children’s speech, and the recordings were normalized for volume. The processed speech recordings were subsequently down-sampled from 44.1 kHz to 16 kHz with 16-bit resolution and saved in standard WAV (PCM) format.

After processing the audio recordings, it was necessary to create a transcription for each of them. The initial speech-to-text transcription was performed automatically using the SARRA[3] ASR system (Lojka 2018). Further adjustments to the transcription were made manually using the Transcriber tool[4] (Barras 2001), where each speech segment was assigned a unique speaker identifier and a time period indicating the start and end of the speaker’s speech. An example of a transcription in the Transcriber tool is shown in Fig. 1.

2.4 Analysis of the database of children’s speech

As shown in Tab. 1, the latest version of the Slovak children’s database consists of 130 children’s TV and radio shows and contains a total of 2,589 speech segments with 303 speakers (127 males and 176 females), amounting to almost five hours of transcribed speech. The average duration of a speech segment is approximately 6.85 seconds. The total number of words in the database is 35,945, of which 3,441 are unique. The age range of child speakers in the database is between 4 and 12 years.

Next, we analyzed the number of out-of-vocabulary (OOV) words. For this calculation, we used the dictionary from the SARRA ASR system, which contains more than 558,000 unique words. Our findings show that the ‘*Oteckovia*’ subcorpus has the highest percentage of OOV words, reaching up to 4.46%. This is quite an interesting result, considering that the ‘*Táraninky*’ and ‘*Rozhlasové leporelo*’ subcorpora contain fully spontaneous speech. These subcorpora have slightly more than 1% of OOV words. Examples of OOV words in individual subcorpora are shown in Tab. 2.

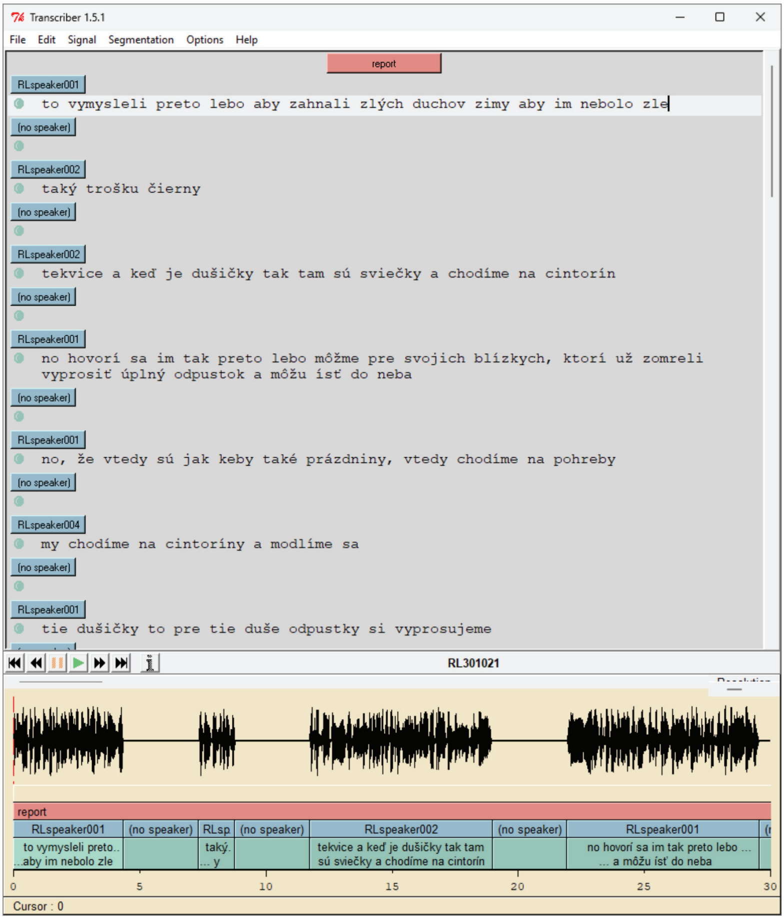
[1] https://www.stvr.sk/archiv

[2] https://www.audacityteam.org/

[3] https://marhula.fei.tuke.sk/sarra/

[4] https://trans.sourceforge.net/

Note that we are still working on transcribing additional sessions and expanding the database. Our goal is to reach at least 30 hours of pure children's speech.



RLspeaker001

to vymysleli preto lebo aby zahnali zlých duchov zimy aby im nebolo zle

(no speaker)

RLspeaker002

taký trošku čierny

(no speaker)

RLspeaker002

tekvice a keď je dušičky tak tam sú sviečky a chodíme na cintorín

(no speaker)

RLspeaker001

no hovorí sa im tak preto lebo môžme pre svojich blízkych, ktorí už zomreli vyprosiť úplný odpustok a môžu ísť do neba

(no speaker)

RLspeaker001

no, že vtedy sú jak keby také prázdniny, vtedy chodíme na pohreby

(no speaker)

RLspeaker004

my chodíme na cintoríny a modlíme sa

(no speaker)

RLspeaker001

tie dušičky to pre tie duše odpustky si vyprosujeme

Fig. 1. The process of transcription of the show '*Rozhlasové leporelo*' in the Transcriber tool

3 MODELING OF CHILDREN'S SPEECH

We used the resulting speech database to train AMs for the task of automatic recognition of children's speech in Slovak. We selected three ASR architectures.

The first architecture is based on the freely available Kaldi ASR engine (Georgescu 2021), which performs speech decoding using Weighted Finite State Transducers (WFST). The baseline triphone AM was trained using a standard procedure based on the extraction of Mel-Frequency Cepstral Coefficients (MFCC) with Cepstral Variance and Mean Normalization (CVMN) to eliminate noise in the extracted speech features. To reduce the dimensionality of the acoustic feature vectors, we applied Linear Discriminant Analysis (LDA) in conjunction with Maximum Likelihood Linear Transform (MLLT) and Speaker Adaptive Training (SAT) to increase the model's robustness to speaker variability. This approach follows the standard GMM-HMM framework for speech recognition, incorporating a trigram language model (LM). Additionally, we tested newer, improved approaches in Kaldi based on 'chain' models, which represent a hybrid DNN-HMM approach. These models utilize a Factorized Time-Delay Neural Network (TDNN-F), which effectively captures long-term temporal dependencies in speech data, as well as an extended version, CNN-TDNN-F, which incorporates a Convolutional Neural Networks as an input layer to extract robust acoustic features. The main advantage of a framework built on the Kaldi recognizer lies in its control over the recognition dictionary, a feature that current architectures based on transformers or self-supervised learning do not provide.

Subcorpus	**'*Oteckovia*'**	**'*Táraninky*'**	**'*Rozhlasové leporelo*'**
Genre	TV family comedy	Kid's TV talk show	Radio sessions for children
Type	narrated scripts	spoken dialogues	spoken dialogues
Years of broadcasting	2018–2019	2020–2023	2021–2023
Age range of speakers	7–12	5–8	4–11
Number of sessions	8	88	34
Number of speakers	6 M / 6 F	43 M / 60 F	78 M / 110 F
Number of utterances	305 (154 M / 151 F)	1,667 (798 M / 869 F)	617 (229 M / 388 F)
Average utterance duration	5.80 seconds	7.20 seconds	6.42 seconds
Number of words	3,970	22,760	9,215
Number of unique words	1,377	4,981	2,521
OOV rate [%]	4.46	1.31	1.36
Total duration (hh:mm:ss)	00:33:13	03:19:45	01:06:00
	04:58:58		

Tab. 1. Statistics on subcorpora of the Slovak children's speech database

'*Oteckovia*'	**'*Táraninky*'**	**'*Rozhlasové leporelo*'**
Ajštaj, baragraft, bratránko, dákam, fotrovi, furt, kakauko, lakťovať, Peťuľka, ocí, potvorko, Prdelákovce, rucpaku, superpes, tato, tatí, tatino, tatuš, tatušó, trampoška, trampošku, vyšokovaný	*barz, Bebík, dáku, dinosaure, elzovský, gaťuše, jaké, jakeby, jakému, jakú, kakaničky, kakauko, našuchne, nesnežná, oblečko, očkatý, odznačky, prečarovať, rozburdať, súrodencovia, tatík, tatíkom, stamaď, sudokmeň, tuna, šlifka, videofotiek, vybáca, zabambuší, zaležaný, zbúraninka, zbúraniny, zlatokopia*	*akrobácie, baletkovú, balzy, docela, drúhe, jaké, jakú, lietatiel, makanie, náťaž, neni, pozauna, snižec, šalený, tehálmi, tote, toten, tuná, zverací*

Tab. 2. Examples of out-of-vocabulary words

To expand the training set, we applied data augmentation, including Speech Perturbation (SP) – modifying the speech rate by factors of 0.9 (slower) and 1.1

(faster) – and Spectral Augmentation (SA). These techniques increased the training set to four times its original size. Other data augmentation techniques, such as perturbation of pitch, volume, tempo, or vocal tract length, did not yield further improvements.

As a final step, after decoding the speech, we also applied post-processing using a LM based on Recurrent Neural Networks (RNN LM).

We chose Whisper[5] (Radford 2023) as the second ASR architecture. It is a closed, End-to-End recognition system implemented as an encoder-decoder Transformer. The decoder is trained to predict the corresponding text caption, interspersed with special tokens that enable the model to perform multilingual ASR, speech translation, and language identification. Whisper's AMs have been trained on a large and diverse multilingual dataset comprising 680,000 hours of audio[6]. Whisper offers several pre-trained models suitable for further fine-tuning. In this research, we evaluated the base, small, medium, and turbo models, with the medium model yielding the best ASR results. Fine-tuning the large model was beyond our computational capabilities.

The third and final architecture we used is Wav2Vec 2.0 (Bayevski 2020), a self-supervised framework for speech representation learning. It is based on a Transformer architecture and learns speech representations by masking parts of raw audio waveforms and predicting them from context. The model is pre-trained on large amounts of unlabeled audio data and can be fine-tuned for ASR with minimal labeled data. There are numerous pre-trained models based on the Wav2Vec 2.0 architecture that are suitable for fine-tuning with Slovak data. Among them, we applied:

- XLS-R-300M[7] (Babu 2022) – a large-scale multilingual ASR model pre-trained on 436,000 hours of unlabeled speech in 128 languages, including Slovak, with 300M parameters;
- MMS-1B-All[8] (Pratap 2024) – a large-scale multilingual ASR model pre-trained on one billion parameters across over 1,000+ languages.

4 EXPERIMENTS AND RESULTS

At the beginning, we divided the database of children's speech in Slovak into a training set and a test set. The test set contained approximately one-third of randomly selected speech segments from the '*Táraninky*' subcorpus, while the remaining data was used for training and validating the models. As a result, the test set included 389 speech segments from 29 speakers (10 males and 19 females), with

[5] https://github.com/openai/whisper

[6] The exact amount of Slovak audio data is not known.

[7] https://huggingface.co/facebook/wav2vec2-xls-r-300m

[8] https://huggingface.co/facebook/mms-1b-all

a total duration of 45 minutes and 57 seconds. The training and validation set was used either to train models from scratch or to fine-tune pre-trained models based on the Whisper or Wav2Vec 2.0 architectures.

We used the standard Word Error Rate (WER) to evaluate the model performance. WER is a common metric for assessing speech recognition performance, calculated as the ratio of the total number of substitutions, deletions, and insertions to the total number of words in the reference transcript.

The results summarized in Tab. 3 show that fine-tuning on children's speech data and applying data augmentation significantly improved ASR performance across all architectures. Kaldi models achieved a WER reduction from 46.10% to 24.19% with CNN-TDNN-F and data augmentation. The fine-tuned Whisper model achieved a WER of 18.29%, outperforming Kaldi. Wav2Vec 2.0 models demonstrated strong performance, with XLS-R-300M fine-tuned on augmented data and a trigram LM achieving a WER of 16.38%. MMS-1B-ALL performed best, reaching the lowest WER of 15.10% when fine-tuned on augmented data with a trigram LM, highlighting the effectiveness of self-supervised learning for child speech recognition.

ASR architecture	**Acoustic model setup**	**WER [%]**
Kaldi	MFCC+CMVN + LDA+MLLT+SAT + trigram LM	**46.10**
	TDNN-F + trigram LM	37.32
	CNN-TDNN-F + trigram LM	37.57
	TDNN-F + RNN LM	35.41
	CNN-TDNN-F + RNN LM	36.05
	TDNN-F + data augmentation (SP+SA) + RNN LM	27.27
	CNN-TDNN-F + data augmentation (SP+SA) + RNN LM	**24.19**
Whisper	medium	44.10
	medium fine-tuned on children's speech data	18.96
	medium fine-tuned on augmented dataset (SP+SA)	**18.29**
Wav2Vec 2.0	XLS-R-300M	41.14
	XLS-R-300M fine-tuned on children's speech data	26.48
	XLS-R-300M fine-tuned on augmented dataset (SP+SA)	25.55
	XLS-R-300M fine-tuned on augmented dataset + trigram LM	**16.38**
	MMS-1B-ALL	34.13
	MMS-1B-ALL fine-tuned on children's speech data	23.86
	MMS-1B-ALL fine-tuned on augmented dataset (SP+SA)	22.45
	MMS-1B-ALL fine-tuned on augmented dataset + trigram LM	**15.10**

Tab. 3. Summary of the results of newly trained or fine-tuned models for children's speech

5 CONCLUSION

Improving children's ASR requires a combination of age-specific adaptation techniques, data augmentation, and self-supervised learning to address data scarcity.

In this research, we compared three different approaches to speech recognition, with self-supervised learning achieving a WER of 15.10%, which is comparable to similar studies (Bhardwaj 2022; Sobti 2024), despite using less than 4 hours of training data.

Future research should focus on expanding child speech corpora, collecting more diverse speech samples from children across various age groups, refining transfer learning techniques, and developing more effective domain adaptation strategies to bridge the performance gap between adult and child ASR. These advancements will enable more accurate and inclusive speech recognition systems for educational, assistive, and interactive speech-oriented applications.

ACKNOWLEDGEMENTS

The research in this paper was partially supported by Ministry of Education, Research, Development and Youth of the Slovak Republic under the research projects KEGA 049TUKE-4/2024 and 041TUKE-4/2025, and by the Slovak Research and Development Agency under the research projects APVV-22-0261 and APVV-22-0414.

References

Babu, A., Wang, Ch., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. In Proc. of INTERSPEECH 2022, Incheon, Korea, pp. 2278–2282.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proc. of NISP 2020, Vancouver BC, Canada, pp. 12449–12460.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communications, Special Issue on Speech Annotation and Corpus Tools, 33(1–2), pp. 5–22.

Batliner, A., Blomberg, M., D'Arcy, Sh., Elenius, D., Giuliani, D., Gerosa, M., Hacker, Ch., Russell, M., Steidl, S., and Wong, M. (2005). The PF_STAR children's speech corpus. In Proc. of INTERSPEECH 2005, Lisbon, Portugal.

Bhardwaj, V., Othman, M.T.B., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., and Hamam, H. (2022). Automatic speech recognition (ASR) systems for children: A systematic literature review. Applied Sciences, 12(9), paper 4419.

Claus, F., Rosales, H. G., Petrick, R., Hain, H.-U., and Hoffmann, R. (2013). A survey about databases of children's speech. In Proc. of INTERSPEECH 2013, Lyon, France.

Eskenazi, M., Mostow, J., and Graff, D. (1997). The CMU kids corpus. LDC97S63. Philadelphia: Linguistic Data Consortium.

Georgescu, A.-L., Pappalardo, A., Cucu, H., and Blott, M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2021(28), pp. 1–30.

Gerosa, M., Giuliani, D., Narayanan, Sh., and Potamianos, A. (2009). A review of ASR technologies for children's speech. In Proc. of WOCCI 2009, Cambridge, MA, USA.

Huber, J. E., and Stathopoulos, E. T. (1999). Formants of children, women, and men: The effects of vocal intensity variation. Journal of Acoustical Society of America, 106(3 Pt 1), pp. 1532–1542.

Lojka, M., Viszlay, P., Staš, J., Hládek, D., and Juhár, J. (2018). Slovak broadcast news speech recognition and transcription system. In: L. Barolli – N. Kryvinska – T. Enokido – M. Takizawa (eds.): Advances in Network-Based Information Systems, LNDECT 22, Springer, Cham, pp. 385–394.

Lu, R., Shahin, M. A., and Ahmed, B. (2022). Improving children's speech recognition by fine-tuning self-supervised adult speech representations. arXiv Preprint. Accessible at: https://arxiv.org/abs/2211.07769.

Patel, T., and Scharenborg, O. (2024). Improving end-to-end models for children's speech recognition. Applied Sciences, 14(6), paper 2353.

Pradhan, S. S., Cole, R. A., and Ward, W. H. (2024). My Science Tutor (MyST) – A large corpus of children's conversational speech. In Proc. of LREC-COLING 2024, Torino, Italia, pp. 12040–12045.

Pleva, M., Ondáš, S., Hládek, D., Juhár, J., and Staš, J. (2019). Building of children speech corpus for improving automatic subtitling services. In Proc. of ROCLING 2019, New Taipei City, Taiwan, pp. 325–333.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevskyi, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25, pp. 1–52.

Radford A., Kim, J. W., Xu, T., Brockman, G., McLeavy, Ch., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proc. of ICML 2023, Honolulu, Hawai, USA, pp. 28492–28518.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. Behavior Research Methods, 51, pp. 1928–1941.

Shivakumar, P. G., and Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis, and recommendations. Computer Speech & Language, 63, paper 101077.

Shobaki, K., Hosom, J.-P., and Cole, R. A. (2000). The OGI kids' speech corpus and recognizers. In Proc. of ICSLP 2000, Beijing, China, pp. 1–4.

Sobti, R., Guleria, K., and Kadyan, V. (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. Multimedia Tools and Applications, 83, pp. 81933–81995.

Yeung, G., and Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. In Proc. of INTERSPEECH 2018, Hyderabad, India, pp. 1661–1665.