Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

AusKidTalk: Using strategic data collection and out-of-domain tools to semi-automate novel corpora annotation

Tünde Szalay[1,2,3,*], *Mostafa Shahin*[1,*], *Tharmakulasingam Sirojan*[1], *Zheng Nan*[1], *Renata Huang*[2,3], *Kirrie Ballard*[2], *Beena Ahmed*[1]

[1]Dept. of Electrical Engineering and Telecommunications, University of New South Wales, Australia
[2]Dept. of Speech Pathology, University of Sydney, Australia
[3]Dept. of Linguistics, Macquarie University, Australia

tuende.szalay@sydney.edu.au

Abstract

Annotating speech corpora for novel populations presents a circular problem: eliminating costly manual transcription requires automatic speech recognition (ASR) tools not yet developed; but developing ASR tools requires annotated speech corpora not available. Manual transcription burden was reduced for AusKidTalk, a novel population due to speaker age and accent, by strategic data collection protocol combined with out-of-domain ASR tools for semi-automatic annotation. The data collection protocol inserted tones and timestamps to automatically segment the recordings. Automatic annotation was conducted by out-of-domain tools for diarisation (NeMo) and orthographic transcription (UNSW ASR). Transcription accuracy with 17% word error rate (WER) for single words and 23% WER for continuous speech allowed for hand-correction instead of transcription, reducing annotation burden. The workflow can be adapted for other corpora and updated with new ASR tools as they become available.

Index Terms: speech corpus, automatic speech recognition, orthographic transcription, Australian English, child speech

1. Introduction

Speech corpora, the large digital collections of transcribed and aligned audio recordings, are crucial resources in speech technology and science, allowing for training automatic speech recognition (ASR) tools and transforming phonetics by revealing variation in speech in new detail through the analysis of large datasets [1, 2]. As accurate manual transcription is labour and cost intensive, corpora are often developed using careful read speech, such as word, sentence, and passage reading tasks, reducing the need for orthographic transcription [3, 4, 5, 6].

Spontaneous speech requires manual or semi-automatic orthographic transcription prior to its use to train ASR models or conduct more detailed analysis (e.g., phonetic transcription) [7, 8, 9, 10]. Manual transcription is less accurate for spontaneous than for scripted speech, and transcription with good reliability and high agreement between transcribers is more time-consuming and thus more costly than quick transcription with lower agreement [11]. ASR tools can reduce the transcription cost by providing automatic orthographic transcription. For example, when building a corpus of spontaneous speech for the high-resource language Italian, automatic transcriptions were generated using freely available services provided by YouTube [10]. However, automatic orthographic transcription had to be reviewed and hand-corrected by a trained researcher to ensure accuracy despite Italian being a high resource language [10].

Developing corpus for a new population presents a circular problem: automatically transcribing a new corpus requires ASR tools not yet developed, but developing ASR tools requires annotated speech corpora from that population. Therefore when building AusKidTalk, the first Australian English (AusE) child speech corpus suitable for developing ASR tools, we designed the data collection procedure to enable a semi-automated annotation workflow that considered the limitations of available speech technologies. We applied an off-the-shelf diarisation tool and developed a custom ASR trained on children speech, as existing ASRs developed for and trained on adult speech perform 2–5 times worse on children's speech despite various feature normalisation and model adaptation techniques [1, 12, 13].

Our results show that considering the availability and limitations of speech technology when designing the data collection protocol can greatly reduce annotation burden. This paper provides guidelines for developing speech corpora for new populations, describing a collection and an automatic annotation workflow that is transferable to different corpora and tasks and can be updated with new technologies as they become available.

1.1. AusKidTalk data description

To date, AusKidTalk has collected data from 556 AusE-speaking children aged 3–12 using three scripted tasks (single word production using picture naming, sentence repetition, non-word repetition) and two semi-spontaneous tasks (story-telling based on picture prompts, emotional speech elicitation) [14].

Task 1 involved presenting 130 pictures, one at a time, to generate 130 individual single or 2-word responses from a child. Target words were designed to capture crucial markers of the AusE accent as well as developmental markers during child language acquisition. Task 3 presented a cartoon video sequence and then prompted the child to tell the story in sentences. The cartoon depicted a green-skinned boy on a skateboard finding a large egg and becoming friends with the green dinosaur that hatched [15]. Having watched the video, children retold the story using their own words based on a series of 13 picture prompts. Picture prompts were presented one-by-one.

Tasks and prompts were presented via a custom software on an Android tablet while speech was recorded onto a PC using a headset microphone and several directional microphones [14]. As the tasks and the prompts appearing on the tablet were not synchronised directly to recorded audio file, the tablet played a 1-second long high-frequency tone at the start of each task and recorded timestamps at the start and end of each task and prompt to assist task- and prompt identification (Sec. 2.1).

[*]These authors contributed equally.

10.21437/Interspeech.2025-539

1.2. Challenges in automatic transcription

As the entire session was recorded, each raw audio file contained five tasks, background noise from the mini-games designed to keep the child engaged, and at least three speakers producing task-related and conversational speech: the child, the pre-recorded model speaker who produced verbal prompts, and the interviewer instructing and aiding the child (e.g., "Can you speak up a bit?", "Good job!"). Occasionally a parent/carer, sibling(s), and/or a project researcher was also present in the room, and their voice may have been audible during conversations.

The audio recordings contained task-related speech as well as varied, spontaneous speech, inherent to children's data. For example, children responded to a picture of a cucumber by saying "zucchini", or by re-telling the story of the entire video during the first picture prompt. There were non-task-related conversations between the child and the interviewer leading the recording session. The combination of unprompted responses, spontaneous conversations, and three distinct speakers resulted in a high volume of non-target speech, increasing the difficulty of automated annotation.

1.3. The goal of the automatic annotation workflow

The goal and scope of the annotation workflow was to 1) identify individual tasks within the audio; 2) identify responses to individual prompts within a task; 3) separate the child's speech from that of the adult(s); and 4) automatically transcribe the child's speech (Fig. 1). The output of the pipeline was a Praat textgrid containing time-aligned automatically generated orthographic transcription of the child's speech recorded with the headset microphone.

2. The automatic speech processing tools

2.1. Task- and prompt-level time alignment

As the first step, the start and end time of Tasks 1 and 3 were identified, as well as the start and end time during which each picture prompt was presented (Fig. 1). To determine the start and end of each task in the audio file, an automatic tone detector was developed using a non-linear binary Support Vector Machine (SVM) with the radial basis function kernel to identify the location of the 1-second long tone. To train the SVM, feature vectors were extracted from 3700 not-tone and 3700 tone frames selected randomly from 10 recordings and spliced with the feature vector of two preceding and two succeeding frames. The SVM classified each 10ms frame of the recording as tone or not-tone. The moving average of the number of detected tones was calculated using a one-second sliding window. Peak points with a moving average above 0.9, i.e., with at least 90% of frames classified as tone within a 1s window, were considered to be tone positions. On a test set of 10 recordings, the classifier achieved 0% false acceptance rate, with not-tone segments never being misidentified as tones. The false rejection rate was approximately 9% with 4/45 tones misidentified as not-tones.

The time duration between every two tones was calculated and compared to the duration between every two timestamps marking the start and the end of a task recorded automatically during the interview. Reference tone-timestamp pairs were identified when the duration between any two tone–sounds was equal to the duration between any two timestamps. Tasks were separated in the raw audio file using reference tone-timestamps. The audio of Tasks 1 and 3 was extracted into separate wav files.

Picture prompts were identified using timestamps recorded by the presentation software during data collection. The start of a prompt interval was recorded when the prompt picture was presented to the child. The end of a prompt interval was recorded when the interviewer pressed the assessment button indicating that the child completed the attempt of the current prompt. Prompt timestamps were mapped onto Praat textgrids with intervals indicating the start and end of each prompt (Fig. 2, top tier).

2.2. Diarisation with NeMo

To separate the child's speech from non-child speech produced by the interviewer, the model speaker, and occasionally by a parent/carer present in the room, the NVIDIA NeMo Speaker Diarisation tool was used as it allowed for being deployed locally and identifying short speaker-intervals, particularly important for a single word production task [16, 17, 18, 19].

NeMo contained five main components. First, the Voice Activity Detection (VAD) component identified the speech intervals within an audio file. The VAD utilised the pretrained MarbleNet model, a Convolutional Neural Network (CNN) designed for speech activity detection [16]. The second component used a multi-scale approach to segment the audio at different scales – typically 2, 1.5, 1, and 0.5 second in length – and extracted speaker embeddings from each [17]. Information from all scales was combined and the final speaker label was determined based on the shortest segment. The third component extracted speaker embeddings from each speech segment using the pre-trained TitaNet-L model [18]. The fourth component clustered the speaker embeddings at each scale of the multi-scale resolutions into estimated speaker clusters by computing a cosine similarity matrix from the embeddings and then applying a spectral clustering algorithm to the similarity matrix and its eigenvalues [19]. Lastly, a multi-scale decoder model assigned speaker labels [17]. The model was trained to weigh the importance of embeddings extracted from segments of varying lengths, facilitating accurate labelling of segments as short as 0.5 seconds.

NeMo's accuracy was tested against the off-the-shelf IBM-Watson model on a sample of the single word production task by four children (age range = 4 – 10 years, mean = 6.75). IBM-Watson Diarisation was accurate with 85%–91% of all target words identified as child speech. As NeMo had comparable accuracy with 92%-100% of target words identified as child speech and was open-source run and controlled locally, NeMo diarisation was selected.

2.3. Automatic word recognition with UNSW ASR

The custom-built University of New South Wales (UNSW) ASR tool was used to orthographically transcribe audio (Fig. 1). The customised UNSW ASR model was developed with an acoustic model trained on American English child speech data and a language model trained on a combination of adult and children's speech transcriptions [20]. The acoustic model was built using a hybrid phoneme-based DNN-HMM architecture, implemented with the Kaldi toolkit [21]. The acoustic model utilised a Factored Time-Delay Neural Network (TDNN-F) architecture trained using lattice-free Maximum Mutual Information (MMI) training criteria [22, 23]. The acoustic model was trained on approximately 380 hours of child data spanning scripted and spontaneous speech collected from around 5,600 children aged 5 to 16 years old from five American speech corpora: 1) the Oregon Graduate Institute Kids' Speech Corpus, 2) the Carnegie Mellon University Kids' Speech Corpus, 3)

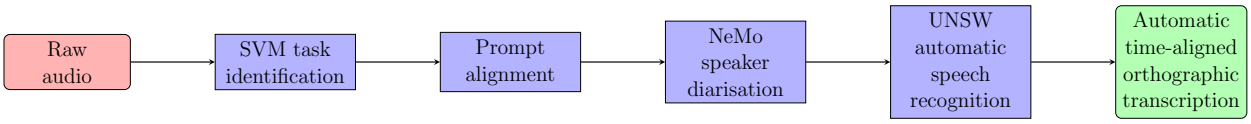

Figure 1: *AusKidTalk workflow to generate automatic orthographic transcription.*

the University of Colorado Kids' Prompted, Read, and Summarized Speech Corpus, 4) the My Science Tutor Children's Speech Corpus, and 5) the Trentino Language Testing Dataset [24, 25, 26, 27, 28]. The original child data were augmented to 2,000 hours using speed perturbation, real Room Impulse Response addition, babble noise, and non-speech noise.

The language model was trained on adult speech transcriptions extracted from TED talks, along with transcriptions from the child speech corpora used for training the acoustic model [29]. To prioritise Task 1 prompts, the language model was interpolated with a specialised model trained on prompts alone. No words were prioritised in Task 3, as the spontaneous story-telling task did not have a fixed vocabulary.

UNSW ASR's accuracy was tested against the off-the-shelf IBM-Watson model on a sample of the single word production task by four children (age range = 4 – 10 years, mean = 6.75) [30]. The accuracy of IBM-Watson was so low as to being practically unusable with a word error rate ranging from 94% to 57% per child [30]. The low ASR accuracy of IBM Watson combined with the preference for an open-source ASR tool run and controlled locally motivated using the UNSW ASR.

2.4. Workflow output

The automated tools of the workflow (i.e., high-frequency tone detection, NeMo, and UNSW ASR) were implemented sequentially (Fig. 1). First, automatic tone detection was used to time-align the audio with the time-stamps, and Task 1 and Task 3 audio were extracted. Extracted audio files with NeMo textgrids containing diarised and time-aligned speech on separate tiers for each speaker were compared to the prompt intervals and fed to the UNSW ASR. When a prompt interval did not overlap with any speaker-intervals identified by NeMo, the entire prompt interval was transcribed by the UNSW ASR system; the recognised words were added to all speaker tiers. For all other prompt intervals, only speaker-segments identified by NeMo were transcribed by the UNSW ASR. The output of the automatic analysis tools was an audio-file segmented on a task-by-task basis, time-aligned with picture prompts, and the speech of all automatically identified speakers transcribed on separate speaker tiers (Fig. 2). Transcription of all speaker tiers were required, as NeMo, while able to separate speakers, could not identify which speaker was the child. At the time of writing, textgrids were generated for 456 Task 1 and 330 Task 3 files.

3. Workflow evaluation

3.1. Testing the workflow

Task- and prompt alignment and diarisation quality were evaluated using auditory and visual observation of the audio- and textgrid files by expert annotators. The annotators compared automatic time-alignment between the audio and the elicitation tasks by listening to the start and end of each audio file to ensure that the audio contained speech data only from the relevant task. The annotators compared the prompts to the speech to evaluate prompt-speech alignment by evaluating whether the child's

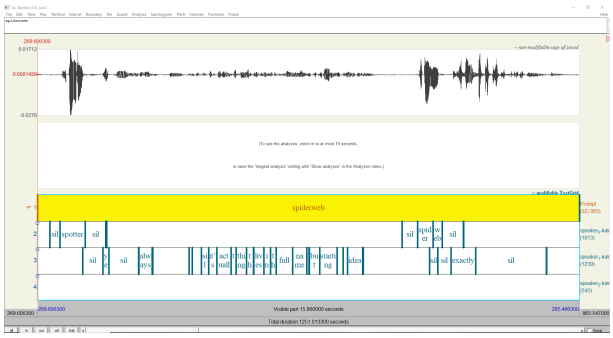

Figure 2: *AusKidTalk workflow output. Top: waveform. Bottom: picture prompt tier followed by time-aligned transcription of three speakers on separate tiers.*

speech matched the expected prompt. To evaluate diarisation quality, the annotators listened to each recording and identified which speaker-tier belonged to the child.

Accuracy of UNSW ASR was evaluated by introducing a hand-correction protocol to correct automatically generated transcription of the child's speech. In Task 1, annotators were instructed to identify intervals that contained the target word matching the picture prompt, accept the automatically generated transcription if it was correct, and edit the transcription if it was incorrect (i.e., it did not match what the child said) [30]. In Task 3, annotators were instructed to identify speaker-turns for each picture prompt that contained only child speech. That is, annotators were instructed to identify intervals during which the child spoke without interruption from the adult. Annotators were instructed to edit the child's turns such that it only contained child speech, accept the automatically generated transcription if it was correct, and edit it if it was incorrect. Automatically generated transcription of adult speech was not hand-corrected. Annotators' were supervised and their work was checked periodically to assist consistency of hand-corrections.

3.2. Task- and prompt alignment performance

At the time of writing, 456 Task 1 and 201 Task 3 files were pre-screened for task- and prompt-alignment. Task and prompt alignment was typically good, with 26 Task 1 (5.7%) and 35 (17%) Task 3 files showing poor alignment.

3.3. Diarisation performance

At the time of writing, 456 Task 1 and 201 Task 3 files were pre-screened for diarisation quality. In Task 1 diarisation, an average of 2.9 speakers (standard deviation = 0.5) were identified out of the expected three (interviewer, child, and pre-recorded model speakers providing verbal prompts). In Task 3, an average of 1.9 speakers (standard deviation = 0.74) were identified per sample, slightly lower than the expected two (interviewer and child - the model speaker did not occur in Task 3). The Task 3 average was qualitatively different from Task 1, as identifying two speakers out of three allowed for separating the child

speaker from the non-child speakers, whereas identifying one out of two speakers indicated failed diarisation.

Reduction in diarisation quality is attributed to Task 3 being shorter than Task 1 (3.5 minutes on average compared to 22 minutes) and to the limited amount of speech produced by the interviewer. Impressionistic observation suggested that the interviewer produced less speech relative to the child in Task 3 than Task 1. While in Task 1, the interviewer frequently provided verbal cues for the children to help them recognise the pictures, verbal prompts in Task 3 were limited to an occasional "What else happened?". As a result, adult and child speech were not always separated well in Task 3.

3.4. Word Error Rate

At the time of writing, 395 Task 1 and 80 Task 3 files were hand-corrected. To evaluate ASR accuracy on comparable Task 1 and Task 3 sets, children with hand-corrected data for both tasks were identified, yielding a set of 71 children. As annotators were instructed to correct child speech transcription only and remove transcription for adult speech, ASR accuracy was evaluated on intervals that contained automatic transcription for child speech only. For Task 3, one child had no turn-intervals containing child speech only due to diarisation errors and interruptions by the adult, therefore one child was excluded from further analysis. In total, UNSW ASR accuracy was evaluated on 2 (tasks) $\times$ 71 (children) - 2 (excluded child) = 140 files.

The test set comprised files from three age groups: 20 children aged 3-5 years (M = 7, F = 13), 22 children aged 6-8 years (M = 13, F = 9), and 28 children aged 9-12 years (M = 18, F = 10). Total audio length was 28.9 hours, consisting of 24.6 hours Task 1 speech (mean duration = 21.1 minutes, standard deviation = 7 minutes) and 4.3 hours Task 3 speech (mean duration = 3.7 minutes, standard deviation = 7 minutes).

Word error rate (WER) was calculated to evaluate UNSW ASR performance. WER was overall lower for Task 1 (16.5%) than for Task 3 (22.7%) (Table 1). Lower WER in Task 1 may be attributed to the UNSW ASR prioritising known target words in Task 1, whereas no words were prioritised in Task 3 containing spontaneous speech with no predefined vocabulary (Sec. 2.3). ASR errors across both tasks are attributed to accent differences between UNSW ASR's American English training data and the AusE accent in the AusKidTalk data. Accent differences are known to adversely impact ASR performance, even between varieties of English [31, 32, 33]. ASR tools developed for American English perform worse on AusE due to phonological and phonetic differences, such as the absence of post-vocalic /ɹ/ in AusE (i.e., the absence of /ɹ/ word-finally or before a consonant, e.g., *car* and *park*) and AusE having a larger vowel inventory containing 18 stressed vowels compared to the 15 stressed vowels of American English [31, 34].

Table 1: *Mean WER (%) by Task, Age group, and Sex (M = male, F = female). Numbers in bold show total WER by Task.*

Age (years)	**Task 1**			**Task 3**		
	M	F	Both sexes	M	F	Both sexes
3–5	29	28	28	29	35	33
6 – 8	14	12	13	23	19	20
9 – 12	13	11	12	20	18	17
All ages	16	17	**16.5**	22	23	**22.7**

WER decreased with age for both sexes and both tasks (Table 1). Younger children are likely to have produced more age-appropriate errors than older children, challenging the ASR. In addition, the American English-speaking children in the training data were older than the youngest AusKidTalk children, increasing age-differences between training and test data. Differences between boys and girls were marginal in Task 1; Task 3 showed larger sex differences with no clear tendency for better performance on either male or female speech. The relatively high WER on AusE-speaking children, in particularly younger children, highlights the need for accent- and age-matched child speech data collection for training ASR tools.

4. Conclusion

In this paper, we demonstrated that understanding how speech technology can be better leveraged at the collection protocol stage can produce a reliable pipeline that semi-automates the data transcription. We validated our semi-automatic annotaion workflow on AusKidTalk, a novel corpus of AusE-speaking children, by collecting data from 556 children completing multiple speech tasks, including a single-word production task and a story-telling task. Automatic task- and prompt-alignment performed well on both tasks as high frequency tones separated tasks during data collection and timestamps of tasks and prompts were recorded by the stimulus-presenting application. NeMo diarisation and UNSW ASR both performed better on Task 1 (single word production task) compared to Task 3 (story-telling task). Better performance of NeMo diarisation in Task 1 is attributed to the high amount of adult speech, while better UNSW ASR performance is attributed to Task 1 having a fixed vocabulary.

UNSW ASR reached sufficient accuracy on both tasks to reduce annotation burden by allowing hand-correction of automatically generated transcriptions. However, hand-correction was still required to achieve high quality aligned transcriptions due to the need to remove adult speech transcription and correct ASR errors. To improve the workflow for the remaining data, the audio file containing all five tasks will be diarised prior to extracting individual tasks to ensure that all speakers provide enough speech for accurate diarisation. Future work will focus on continuing hand-correcting Task 1 and 3 data and on fine-tuning the UNSW ASR tool. The resulting corpus will be suitable for phonetic analysis of AusE child speech as well as for ASR development.

The corpus was co-designed by engineers developing data collection methods with high-frequency tones facilitating semi-automated annotation methods to reduce annotator burden, and by phoneticians and speech language pathologists designing speech tasks suitable for children. The AusKidTalk annotation workflow concatenated multiple automatic tools in a step-by-step manner (Fig. 1). As downside of the concatenation, failure of an initial step affected the performance of the following steps. For example, when the first step, i.e., task identification, failed either due to incorrect high-frequency tone detection or due to the tones not being played during data collection, the relevant tasks were not extracted and the following steps were not completed. The workflow cannot be deployed on a few recordings collected using a pencil-and-paper protocol due to tablet failure. The benefit of using a series of automatic speech processing tools is easy substitution of new tools if/when state-of-the-art in-domain diarisation or ASR tools become available. The described workflow is transferable to other corpora and speech tasks and highlights the need for co-designing corpora by researchers in speech science and technology.

5. Acknowledgement

We would like to thank our participants without whom this project would not have been possible. This project was supported by the Australian Research Council LE190100187 and FT180100462 grants, as well as the University of New South Wales, The University of Sydney, Western Sydney University, Macquarie University and The University of Melbourne. This project was approved by The University of New South Wales Human Research Ethics Approval HC190320.

6. References

[1] R. Sobti, K. Guleria, and V. Kadyan, "Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges," *Multimedia Tools and Applications*, pp. 1–63, 2024.

[2] M. Y. Liberman, "Corpus phonetics," *Annual Review of Linguistics*, vol. 5, no. 1, pp. 91–107, 2019.

[3] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[4] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner *et al.*, "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box," in *Interspeech*, 2011, pp. 841–844.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[6] A. A. Kressner, K. M. Jensen-Rico, J. Kizach, B. K. L. Man, A. K. Pedersen, L. Bramsløw, L. B. Hansen, L. W. Balling, B. Kirkwood, and T. May, "A corpus of audio-visual recordings of linguistically balanced, Danish sentences for speech-in-noise experiments," *Speech Communication*, vol. 165, p. 103141, 2024.

[7] N. Grønnum, "A Danish phonetically annotated spontaneous speech corpus (DanPASS)," *Speech Communication*, vol. 51, no. 7, pp. 594–603, 2009.

[8] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[9] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Speech Communication*, vol. 94, pp. 62–74, 2017.

[10] D. Mereu and A. Vietti, "Dialogic ItAlian: the creation of a corpus of Italian spontaneous speech," *Speech Communication*, vol. 130, pp. 1–14, 2021.

[11] M. L. Glenn, S. M. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription methods for consistency, volume and efficiency." in *LREC*, 2010.

[12] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (ASR) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.

[13] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.

[14] B. Ahmed, K. Ballard, D. Burnham, T. Sirojan, H. Mehmood, D. Estival, E. Baker, F. Cox, J. Arciuli, T. Benders *et al.*, "AusKidTalk: an auditory-visual corpus of 3- to 12-year-old Australian children's speech," in *Interspeech*, 2021, pp. 3680–3684.

[15] Doggy dog. [Online]. Available: https://www.youtube.com/watch?v=DTfv8y05fj8

[16] F. Jia, S. Majumdar, and B. Ginsburg, "MarbleNet: Deep 1D time-channel separable convolutional neural network for voice activity detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6818–6822.

[17] T. J. Park, N. R. Koluguri, J. Balam, and B. Ginsburg, "Multi-scale speaker diarization with dynamic scale weighting," in *Interspeech*, 2022, pp. 5080–5084.

[18] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[19] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.

[20] M. A. Shahin, R. Lu, J. Epps, and B. Ahmed, "UNSW system description for the shared task on automatic speech recognition for non-native children's speech." in *Interspeech*, 2020, pp. 265–268.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[22] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech*, 2018, pp. 3743–3747.

[23] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755.

[24] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000, pp. 564–567.

[25] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids corpus ldc97s63," Web Download, Philadelphia, 1997.

[26] R. Cole and B. Pellom, "University of Colorado read and summarized story corpus," in *Technical Report TR-CSLR-2006-03*. University of Colorado, 2006.

[27] W. Ward, R. Cole, and S. Pradhan, "My science tutor and the MyST corpus," *Boulder Learning Inc*, 2019.

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

[29] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer*. Springer, Cham, 2018, pp. 198–208.

[30] T. Szalay, L. Ratko, M. Shahin, T. Sirojan, K. Ballard, F. Cox, and B. Ahmed, "A semi-automatic workflow for orthographic transcription of a novel speech corpus: A case study of AusKidTalk," in *Proc. of 18th SST*, R. Billington, Ed. Canberra: ASSTA, 2022.

[31] T. Szalay, M. Shahin, B. Ahmed, and K. Ballard, "Knowledge of accent differences can predict speech recognition errors," in *Interspeech*, 2022, pp. 1372–1376.

[32] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," *Speech Communication*, vol. 140, pp. 50–70, 2022.

[33] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *Interspeech*, 2017, pp. 934–938.

[34] T. Szalay, M. Shahin, K. Ballard, and B. Ahmed, "Training forced aligners on (mis) matched data: the effect of dialect and age," in *Proc. of 18th SST*, R. Billington, Ed. Canberra: ASSTA, 2022.