Language Resources and Evaluation (2026) 60:53
https://doi.org/10.1007/s10579-026-09929-5

ORIGINAL PAPER

AusKidTalk: developing an orthographic annotation workflow for a speech corpus of Australian English-speaking children

Tünde Szalay[1,2,3] · Mostafa Shahin[2] · Tharamkulasingam Sirojan[2] · Zheng Nan[2] · Renata Huang[1,3] · Joanne Arciuli[4] · Elise Baker[5,6,7] · Felicity Cox[3] · Kirrie Ballard[1] · Beena Ahmed[2]

Received: 3 June 2025 / Accepted: 10 June 2026

Abstract

Child speech corpora are limited in number and scope, with none available for Australian English (AusE) until now, primarily due to orthographic transcription costs. Therefore, we developed AusKidTalk, the first AusE child speech corpus, a novel population due to speaker age and accent. Annotating AusKidTalk presented a circular problem: eliminating costly manual transcription required automatic speech recognition (ASR) tools not yet developed; but developing ASR tools required annotated speech corpora not available. This paper demonstrates how orthographic transcription burden was reduced in AusKidTalk via strategic data collection combined with out-of-domain ASR tools for automatic annotation augmented by manual correction. 620 children completed a single word production task, orthographic transcriptions were generated for 454 (73%) using the semi-automatic AusKidTalk pipeline and corrected manually for 394 using a custom Praat interface. An additional 58 children were transcribed without ASR assistance for comparison, yielding a total of 461 (74%) transcriptions. Workflow efficiency was evaluated on 380 children, producing a total of 136.6 h of speech. Annotation burden was reduced by (i) automatic orthographic transcription with 20% word error rate, removing the need to transcribe speech from scratch, and by (ii) the Praat interface allowing 136.6 h of speech to be corrected by listening to 18.5 h. Annotator reliability was good, with annotators achieving high similarity on 43/49 ground truth files. Correction time for one child was typically 1.5 h compared to the 4 h for transcription unassisted by ASR. The workflow can be adapted for other corpora and updated with new ASR tools.

Keywords Child speech · Corpus building · Speech database · Automatic speech recognition · Speaker diarisation · Assisted speech transcription

Extended author information available on the last page of the article

Published online: 06 July 2026

1 Introduction

Speech corpora, the large digital collections of transcribed and aligned speech recordings, are crucial resources in speech science and technology (Sobti et al., 2024; Liberman, 2019). In phonetics, corpora allow for the study of variation in speech using datasets several orders of magnitude larger than what has previously been available for researchers, revealing phonetic, phonological, social, and communicative variation in speech in new detail (Liberman, 2019). In speech technology, large corpora are required for developing new tools, such as automatic speech recognition (ASR) systems or phonetic aligners, through training machine learning models (Sobti et al., 2024). Such tools perform poorly when the training corpora differ from the new speech data on which they are applied: ASR models trained on typical speech do not generalise well to disordered speech (Schultz et al., 2021), models trained on standard American English do not generalise to other accents, such as African American Vernacular English (Wassink et al., 2022) or Australian English (Szalay et al., 2022b), and models trained on adult speech may perform up to five times worse on child speech (Potamianos and Narayanan, 2003; Shivakumar and Georgiou, 2020; Sobti et al., 2024). These results collectively demonstrate that speech and language variation limits the generalisability of speech technologies, highlighting the need for in-domain corpora, i.e., training data that match the speech data on which the new tool will be applied.

However, the quantity, quality, and availability of in-domain corpora vary between languages and varieties (Besacier et al., 2014). There are multiple corpora available for English-speaking adults, covering a wide range of speech registers, including read speech, TV broadcasts, and informal, spontaneous speech e.g.,, (Garofolo et al., 1993; Pitt et al., 2005; Burnham et al., 2011; Panayotov et al., 2015; Sprugnoli et al., 2017). For adult speakers of other languages, such as Icelandic, Catalan, and Finnish, parliamentary speeches, together with their transcriptions provide transcribed speech data for ASR development (Helgadóttir et al., 2017; Kulebi et al., 2022; Virkkunen et al., 2023). In contrast, child corpora are limited in number and scope due to the difficulties associated with child speech data collection and transcription (Sobti et al., 2024). This paper demonstrates how the lack of child speech corpora was remedied by developing AusKidTalk, the first corpus for Australian English-speaking children and provides guidelines for developing speech corpora for new populations, describing an automatic annotation workflow augmented by hand correction.

1.1 Child corpora

Child corpora are limited in number (Chen et al., 2016; Radha et al., 2024; Sobti et al., 2024). In 2016, only 15 children's speech corpora existed worldwide (Chen et al., 2016). Of these, only three had a sufficient size for ASR development with more than three hundred speakers; the largest had 200 h of speech (Chen et al., 2016). By 2024, the number of child speech corpora used for ASR and automatic speaker identification and verification had increased to 14 English corpora, covering British, American, and non-native Englishes, and to 11 corpora in languages other than English (Radha et al., 2024; Sobti et al., 2024).

Yet, child corpora are limited in the scope of speech tasks and speaker age. A key challenge of developing child speech corpora is speech data collection (Sobti et al., 2024). Read speech tasks using written prompts are the most common modality for speech data collection, as they allow control over the phonetic content and do not require speech transcription, only verification (Besacier et al., 2014; Sprugnoli et al., 2017; Sobti et al., 2024). Therefore, many child corpora have only collected speech data from literate children using written prompts (e.g., CMU Kid, OGI, PF_STAR, TARBALL, UC Kid, SingaKids-Mandarin) (Eskenazi et al., 1997; Shobaki et al., 2000; Batliner et al., 2005; Kazemzadeh et al., 2005; Cole and Pellom, 2006; Chen et al., 2016). In these corpora, prompts were presented predominantly through text, with optional audio and/or picture prompt to make the task more engaging (Shobaki et al., 2000; Kazemzadeh et al., 2005; Chen et al., 2016). Thus, the written prompt served as transcription, removing the need for costly orthographic transcription and replacing it with speech verification (Shobaki et al., 2000; Kazemzadeh et al., 2005). For example, children's speech was verified as "correct" during data collection when their speech matched the prompt in the OGI Kids' Speech Corpus and the TARBALL dataset(Shobaki et al., 2000; Kazemzadeh et al., 2005).

The use of written prompts, however, limits the lower age of the children in the corpora to literate children, typically aged 5 and above (Sobti et al., 2024; Bhardwaj et al., 2022). These corpora do not capture developmental speech and language differences between younger (3–5) and older (6–9, 10 –12) children, such as differences in the production of late-acquired sounds, consonant clusters, and vocabulary, although some of these differences, such as consonant cluster simplification lead to increased ASR errors (Hämäläinen et al., 2014). Thus, the availability of ASR technologies for younger children is limited, despite children being potential beneficiaries of compelling applications of speech and language technology such as remote speech therapy tools, interactive reading tutors, pronunciation coaching, and educational games. The Child Language Data Exchange System (CHILDES) project focuses on children below 5 (Kempe et al., 2024); however, it is of limited use for child ASR development. It contains speech data from multiple projects, yielding a series of small and inconsistent datasets across various languages, with less than half of the database containing publicly available audio data, potentially to protect participants' privacy (Kempe et al., 2024).

1.2 Transcribing child speech corpora

When written prompts are not suitable for child speech elicitation, e.g., due to the child's age or due to the corpus requiring spontaneous speech, picture prompts yield structured speech, and playroom-like setups are used to elicit spontaneous speech e.g.,, (Gao et al., 2012; Pérez-Espinosa et al., 2020; Rumberg et al., 2022). For such non-read tasks, orthographic transcription is required prior to more detailed phonetic, syntactic, or lexical analysis and to facilitate the development of ASR tools. When orthographic transcription is conducted manually, often multiple annotators work on the same data in an iterative consensus procedure e.g.,, (Gao et al., 2012; Rumberg et al., 2022). For example, German continuous child speech was transcribed by three annotators consecutively, firstly and secondly by a speech-language pathologist (stu-

dents or graduate), then by a professional transcribing company, with differences in transcription discussed until a consensus was reached (Rumberg et al., 2022). Mandarin spontaneous child speech from children as young as one year old was transcribed and first reviewed by a team of transcribers, then by a senior transcriber (Gao et al., 2012). In case of minor disagreements between the initial and the senior transcribers, the senior transcriber corrected the errors, while for major disagreements, speech was transcribed again by the first team. The Spanish IESC-Child corpus consisting of 174 children employed two transcribes, but did not report an iterative consensus procedure (Pérez-Espinosa et al., 2020).

Transcription time of child corpora is not always reported, but transcription-time of adult speech may be 10-fold of speech time depending on the complexity of speech to be transcribed (Bazillon et al., 2008). Child speech is expected to be more time-consuming due to reduced intelligibility and increased ambiguity caused by developmental speech errors, contributing to the potentially high cost and time demand of manual transcription. For instance, IESC-Child, consisting of 174 children and 38,147 words, was transcribed in two months by two annotators; however, work hours were not reported (Pérez-Espinosa et al., 2020). The need for multiple trained annotators and agencies can further increase costs.

Using existing ASR tools to generate automatic orthographic transcription and augmenting it with manual correction is known to reduce the time demand of speech transcription compared to manual transcription for adult corpora (Garofolo et al., 1993; Bazillon et al., 2008; Liberman, 2019). Automatic alignment augmented by hand-correction was first used at a large scale for the creation of the TIMIT corpus, followed by other corpora, such as the French EPAC corpus and a multilingual dataset of television speech (Garofolo et al., 1993; Liberman, 2019; Bazillon et al., 2008; Sprugnoli et al., 2017). In the EPAC corpus, hand-correcting two hours of scripted adult speech was completed in 8.5 h compared to the 17.5 h required for manual transcription (Bazillon et al., 2008). Using automatic transcription for child speech in the CHILDES project was less successful: while the Batchalign ASR tool was developed for generating automatic transcriptions that meet the requirements of CHILDES format, they were only formally evaluated on adult speech (Liu et al., 2023). Batchalign reached 95% accuracy on adult speech, and its performance was evaluated as "good" for school-aged children and as "moderate" for school aged-children with mild speech disorders without reporting word error rate (Liu et al., 2023). However, ASR accuracy for young children's speech remained an issue (Kempe et al., 2024).

The poor performance of ASR tools on child speech creates a circular problem: child speech corpus transcription is limited by the lack of child ASR and child ASR development is limited by the lack of child speech corpora. Despite the increase in child speech corpora, the majority of children's ASR tools have been developed using the same six datasets (WSJCAM0, PF_Star, TIDIGIT, ChildIt, TIMIT, and CMU Kid), out of which only three (PF_Star, ChildIt, and CMU Kid) contained exclusively child speech in American- and British English and Italian (Bhardwaj et al., 2022; Radha et al., 2024). Two out of the six most popular corpora for child ASR development contained adult speech and one contained both adult and child speech (Bhardwaj et al., 2022). Other child corpora for American English are available, such as My Science Tutor, CU and OGI, (Ward et al., 2019; Cole and Pellom, 2006; Shobaki et

al., 2000); however, they were not among the frequently used corpora for child ASR development (Bhardwaj et al., 2022). OGI is potentially used less frequently due to its limited vocabulary, while My Science Tutor is limited to American English-speaking children from 3rd to 5th grade.

1.3 Aims

Until now, there has been no sizeable or available speech corpus for Australian English (AusE)-speaking children. For AusE, small datasets have been collected using a single task (e.g., picture naming), with participant numbers ranging from 20 to 255 speakers, often to answer specific research questions on lexical stress acquisition or sound change e.g.,, (Millasseau et al., 2021; Gibson et al., 2023; Arciuli et al., 2019). In the Future Proofing Study, large scale data were collected from over 1000 adolescents aged 12 or above using a variety of speech tasks, but speech tasks were not transcribed systematically (Werner-Seidler et al., 2023). As a result, there is currently no annotated child speech corpus for ASR development for AusE, limiting the availability of ASR tools for this population. Tools developed for other varieties, such as American English, perform worse on AusE due to accent differences; while ASR tools developed for AusE-speaking adults perform worse on children due to age-related linguistic differences (Szalay et al., 2022b, 2022c).

Our aim was to remedy the lack of an AusE child speech corpus and break the circular problem of novel corpus development. Therefore, when building AusKidTalk, the first AusE child speech corpus suitable for developing ASR tools, we designed a non-read data collection procedure to enable a semi-automated annotation workflow that considered the limitations of available speech technologies. This paper provides guidelines for developing speech corpora for new populations, describing a semi-automatic annotation workflow augmented with manual correction for a single word production task that is transferable to different corpora and tasks and can be updated with new technologies as they become available.

2 AusKidTalk: data description

The AusKidTalk project was approved by The University of New South Wales Human Research Ethics Committee (HC190320).

2.1 Participants

The corpus contains audiovisual data from 620 children aged 3–12 years (Table 1). The children are native speakers of AusE with at least one parent having com-

Table 1 Number of children in the AusKidTalk corpus by speaker age and gender

Total number of participants in bold

Age (years)	Girls	Boys	Total
3–5	103	100	203
6–8	103	114	217
9–12	93	107	200
Total	299	321	**620**

pleted their schooling in Australia. Participants are children with typically developing speech (545 children, 92%), as well as children with speech sound disorder (21 children, 3.5%) or autism spectrum disorder (25 children, 4.22%) or both (1 child, 0.16%). Participants received $50 for their time.

2.2 Material and procedure

Participants completed a total of five tasks to elicit speech: (1) single-word production with picture naming; (2) sentence repetition with audio prompt; (3) narrative production with video and picture prompts (4) emotion elicitation with video prompt; and (5) non-word repetition with audio prompt. As the focus of the current paper is the annotation of the single-word production task, hereafter referred to as Task(1), we focus on description and design considerations for Task(1). For details on the other tasks, see Ahmed et al. (2021).

Task(1) was a single-word production task elicited via picture prompts using The Speech Test for Australian Children (STAC) (Baker et al., 2019). The STAC was designed to gather recordings of each child producing target words spontaneously via picture naming or via imitation, if needed, using a standard pre-recorded spoken prompt. Each child was prompted to produce 117 single words (115 STAC items and two practice words), numbers from 1 to 10, and 12 two-word counting phrases (e.g. *one egg, two eggs, three eggs, four eggs, one cup*), giving a total of 139 targets. The STAC collectively sampled all AusE consonants at least once in phonotactically permissible word-initial, -medial and -final positions, and in a range of word-initial and -final consonant clusters. The STAC also sampled all AusE vowels in varied phonetic contexts. Words varied in syllable length (68 monosyllables, 19 disyllables, 30 polysyllables) to sample variation in polysyllabic word production and in lexical stress pattern (trochaic and iambic) to allow for analysis of age-dependent acquisition of weak-strong syllables relative to strong-weak patterns in English (Ballard et al., 2010, 2012; Arciuli and Ballard, 2017; Arciuli et al., 2019). Potential examples of ongoing sound changes were also included. For instance, two targets contained an /ɹ/-cluster which may be the site of pre-/ɹ/ /s/-retraction (Stevens and Harrington, 2016). There were 12 words with vowels before dark coda /l/ potentially undergoing pre-/l/ vowel change (Gibson et al., 2023; Szalay et al., 2023).

Speech tasks were presented by an interviewer via a tablet. The interviewer controlled which task was presented to the child, whether to repeat, skip or go back to a stimulus within the task, and whether to instruct the child to produce the target or to play the pre-recorded audio prompt. If a child was losing interest, the interviewer could choose to pause and present the child with one of several custom, non-speaking games integrated into the stimulus presentation software to re-engage them; or if a child could not continue, the interviewer could exit that activity.

2.3 Recording equipment

Existing equipment from the AusTalk study – The Big Australian Speech Corpus: An audio-visual speech corpus of AusE-speaking adults (Burnham et al., 2011; Wagner et al., 2011) – was upgraded and made child friendly. Audio was recorded via

a total of four microphones: a DPA4088 cardioid headset microphone worn by the child, a Shure Microflex far-field microphone, and two Behringer Studio Condenser Microphones C-2 placed on the table in front of the child. The sampling frequency of the primary headset microphone was 44100 Hz. All microphones were fed into an M-Audio Fast Track Ultra 8R audio mixer, connected to a main recording computer. For details of recording setup, see Ahmed et al. (2021).

To facilitate automated speech segmentation and annotation of the recordings, timestamps were generated each time the interviewer interacted with their tablet application. This included timestamps for each time a task started and ended, each time a stimulus was presented to the child (both onset and offset times), and each time the child responded to or re-attempted a stimulus. In addition, the tablet played a high frequency tone at the start of each of the five tasks. This tone was captured by the recording device and used to synchronise the tablet-generated timestamps with the audio recordings. Recordings and timestamps were saved in the local drive and on an external hard drive, then backed up automatically on the project-specific server established at The University of New South Wales (UNSW), as well as in the UNSW Data Archive.

2.4 Challenges of AusKidTalk transcription

As the entire session was recorded, each raw audio file contained five tasks, background noise from the mini-games, 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?", "Very good!"). Occasionally a parent/carer, siblings, 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 cicada by saying "cricket", or with a giggle to a picture of a bellybutton. 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.

3 The AusKidTalk workflow

We developed a semi-automatic method to transcribe and annotate task-related speech in Task(1). This workflow provides orthographic transcription of all 139 targets and locates their start and end times in the audio file for each child. The workflow concatenated a series of custom-built and out-of-box automated tools and augmented them using human annotation (Fig. 1). The output of our workflow is a time-aligned Praat textgrid that contains intervals for the Task(1) target words only and an easily searchable csv file listing target words and responses with their start and end times. Workflow considerations and preliminary results were reported in Szalay et al. (2022a).

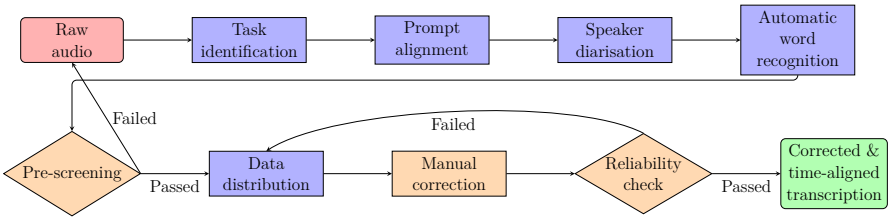

Fig. 1 Workflow outline. Red: start point. Green: end point. Blue: automated process. Yellow: manual process. Diamond: decision point. Sec. 3.1.1: Task identification and prompt alignment; Sec. 3.1.2: Speaker diarisation; Sec. 3.1.3: Word recognition; Sec. 3.2: Pre-screening and data distribution; Sec. 3.3: Manual correction; Sec. 3.4: Reliability check

3.1 Automatic pre-processing tools

Three automatic speech processing tools were used, two of which were custom-built for AusKidTalk. For technical details of these tools, see Szalay et al. (2025); the code is available via Author Mostafa Shahin's GitHub.[1]

3.1.1 Task- and prompt-level time alignment

As the first step, the start and end time of Task(1) was 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 for AusKidTalk (Szalay et al., 2025). On a test set of 10 recordings, the tone detector 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 tone–sounds 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 (Fig. 2). 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 audio file using reference tone-timestamps and Task(1) audio was extracted into a separate wav file.

Prompts were identified using time-stamps recorded by the prompt presentation software (Fig. 2). The start of a prompt interval was recorded at the time 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. Praat textgrids were generated with intervals indicating the start and end of each prompt using the prompt timestamps recorded by the presentation software (Fig. 2).

[1] https://github.com/mostafashahin/AusKidTalkV3.

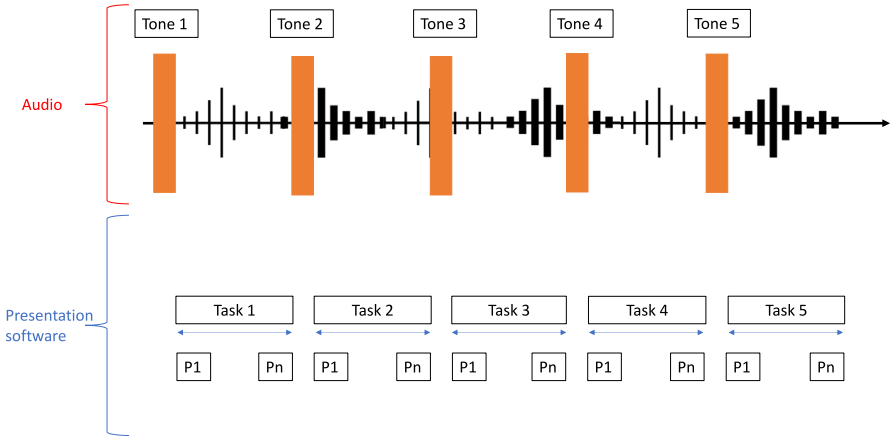

Fig. 2 Time aligning the audio with task and prompt (P) intervals. Five tasks are separated by five tones, detected automatically (top). For each task, presentation time of tasks and prompts were recorded by the presentation software (bottom)

3.1.2 Diarisation with NeMo

To separate the child's speech from non-child speech, the commercially available IBM-Watson diarisation tool and the freely available NVIDIA NeMo Speaker Diarisation tool were tested (Szalay et al., 2022a, 2025). While IBM-Watson reached good accuracy, correctly separating the child's speech from that of the adults in a set of five recordings (Szalay et al., 2022a), lack of control over the tool (e.g., changes in the diarisation system during the project's timeline) and the need to upload the data to an on-line server for processing made the use of IBM-Watson unfeasible. Therefore, NVIDIA NeMo was selected, as it can be deployed locally and reached comparable accuracy. The output of NeMo diarisation was a textgrid with one to four speaker-tiers, with each tier showing intervals for the start and end of a given speaker's speech.

3.1.3 Automatic word recognition with UNSW ASR

The custom-built UNSW ASR tool was used to orthographically transcribe audio due to the lack of suitable off-the-shelf tools (Shahin et al., 2020). IBM-Watson ASR was tested on a sample of five children for orthographic transcription; however, accuracy was so poor so that it was practically unusable with a word error rate ranging from 57 to 94% per child (Szalay et al., 2022a). Therefore, the customized 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, tailored to prioritise Task(1) prompts (Shahin et al., 2020).

UNSW ASR's acoustic model was built using the Kaldi toolkit (Povey et al., 2011, 2016) and trained on 380 h of original child speech data (Shahin et al., 2020). Child speech data were sourced from five American speech corpora spanning scripted and spontaneous speech from around 5600 children aged 5 to 16 years: the Oregon Grad-

uate Institute (OGI) Kids' Speech Corpus (Shobaki et al., 2000), the Carnegie Mellon University (CMU) Kids' Speech Corpus (Eskenazi et al., 1997), the University of Colorado (CU) Kids' Prompted, Read, and Summarized Speech Corpus (Cole and Pellom, 2006), the My Science Tutor (MyST) Children's Speech Corpus (Ward et al., 2019), and the Trentino Language Testing (TLT-School) Dataset (Gretter et al., 2020). The original child data was augmented to 2000 h using speed perturbation, real Room Impulse Response addition, babble noise, and non-speech noise (Shahin et al., 2020).

UNSW ASR's language model was trained on transcriptions of adult speech extracted from TED talks, along with transcriptions from the children's speech corpora used for training the acoustic model (Rousseau et al., 2014; Shahin et al., 2020). To prioritise Task(1) prompts, the language model was interpolated with a specialized model trained solely on the prompts (Shahin et al., 2020).

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 the Task(1) audio were extracted. Task(1) audio files with NeMo textgrids containing diarised and time-aligned speech were compared to the prompt intervals and fed to the UNSW ASR. For every prompt interval, the speaker-segments identified by NeMo were transcribed by the UNSW ASR. When NeMo did not identify any speaker-segment for a prompt interval, the entire prompt interval was transcribed by the UNSW ASR system and the recognised words were added to all speaker tiers.The output of the automatic analysis tools was Task(1) audio extracted and time-aligned with picture prompts, and the speech of all automatically identified speakers transcribed on separate speaker tiers (Fig. 3). Transcription of all speaker tiers was required, as NeMo,

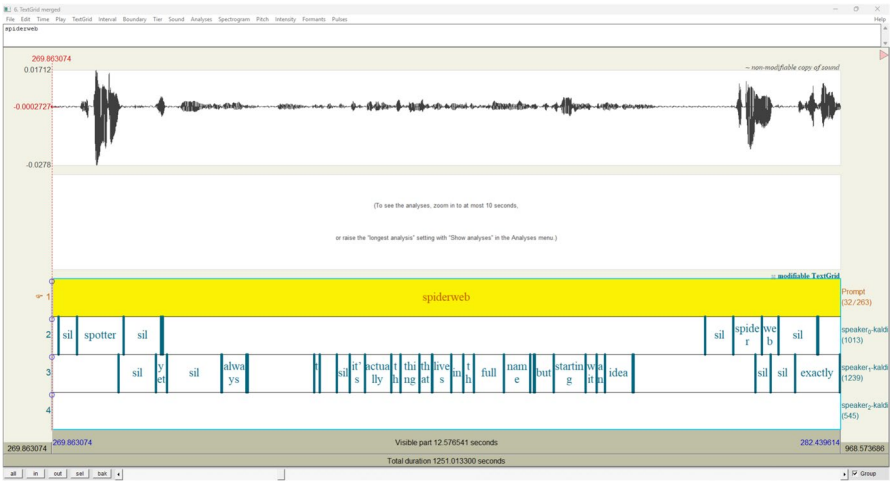

Fig. 3 Praat textgrid generated by automatic time-alignment, diarisation, and word recognition in Task 1. Tier 1: the picture prompt for *spiderweb* was displayed. Tier 2: transcription of the child's speech with *spider* transcribed as *spotter* and *spiderweb* as *spider web*. Tier 3: transcription of the interviewer's speech. Tier 4: model's speaker tier. The model speaker's voice was not played for this prompt

while able to separate speakers, could not identify which speaker was the child. For further details on the automatic pre-processing tools, see Szalay et al. (2025).

3.2 Pre-screening and distributing the data for manual correction

The extracted Task(1) audio files, together with the automatically generated textgrids were pre-screened by expert annotators for audio quality, and errors in time-alignment and diarisation (Fig. 4). Audio- and textgrid pairs that passed the preliminary quality check were distributed to research assistants to correct transcription and time-alignment of target speech.

3.2.1 Pre-screening the data

The data were pre-screened to exclude files with poor audio-quality, incorrect task-identification and prompt-level alignment (i.e., failed high-frequency tone detection), and poor speaker-diarisation from hand-correction. Pre-screening was necessary to reduce manual annotation time: data with poor quality were more time-consuming and thus more expensive to hand-correct. To correct data from as many speakers as possible, only files with high-quality automatic annotations were hand-corrected.

Pre-screening was carried out using auditory and visual observation of the audio- and textgrid files by expert annotators using Praat (Boersma and Weenink, 2023). First, the annotators were prompted to evaluate the quality of the audio by looking at

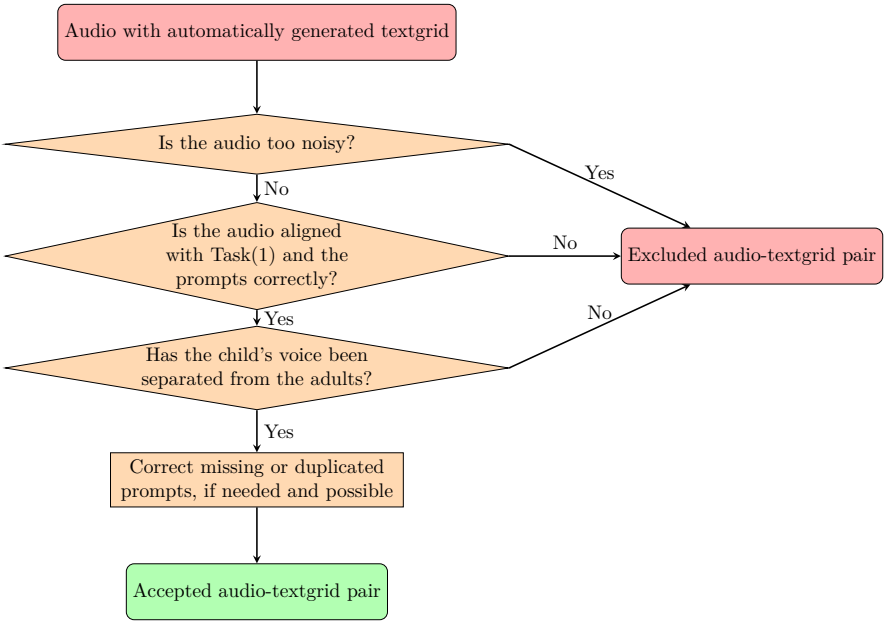

Fig. 4 Workflow used to pre-screen the automatically generated textgrids. Red: start point and end point for excluded files. Green: end point for files passing pre-screening. Yellow: manual process. Diamond: decision point

the spectrogram, using default spectrogram settings, and listening to the audio. Files deemed to be too noisy to easily identify word-boundaries were excluded from hand-correction (Fig. 4, for number of files excluded see Sect. 4).

Second, the annotators compared automatic time-alignment between the audio and the elicitation tasks, and the audio and the prompts by listening to the speech and comparing it to the prompt tier. The annotators were instructed to listen to the start and end of each audio file to ensure that the audio contained speech data only from Task(1). In addition, the annotators compared the prompts to the speech to evaluate prompt-speech alignment. For example, the alignment was correct when the prompt was *watermelon* and the child talked about watermelons. Files with incorrect task- and prompt-alignment were excluded from hand-correction (Fig. 4).

Third, diarisation quality was evaluated by auditory comparison of speech and speaker-tiers. The annotators were instructed to accept diarisation in which the child's speech was mapped onto a single tier with little to no speech from other speakers irrespective of successful separation of the other two speakers (interviewer and model speaker). That is, as long as the child's speech was successfully separated from other speakers, diarisation was accepted as correct and the child's tier was identified.

As the fourth and last step in the pre-screening process, the annotators reviewed the quality of prompt-labels by identifying duplicated or missing prompts. For missing prompts, the annotators were instructed to find the time at which the missing prompt was expected to occur based on presentation order, and insert the missing prompt interval based on listening to the audio. For example, the prompt *cicada* was always presented between *crocodile* and *caterpillar* during data collection. If *cicada* was flagged as missing from the prompt tier, the annotator listened to the intervals for *caterpillar* and *crocodile*. If the audio indicated that *cicada* was displayed (e.g., the interviewer said "Well, it is a kind of an insect, it makes a very loud noise" and the child responded by saying "cicada"), the annotator inserted a missing prompt interval for *cicada* at the relevant timepoint. If a prompt occurred multiple times, the ones occurring at the incorrect time were removed.

To streamline pre-screening, a Praat interface was developed consisting of a series of four evaluation points with simple instructions presented in four consecutive Praat pop-up windows along with the waveform, spectrogram, and textgrids showing the prompts on one tier and the three speakers on three separate tiers. To reduce pre-screening time, the evaluation points were ordered in a hierarchical manner, from least time-consuming to answer to most time-consuming (Fig. 4). That is, if an audio file was evaluated as too noisy in the first step, the file was immediately excluded from analysis and the following questions regarding time-alignment, diarisation quality, and incorrect prompt intervals were not answered. Similarly, if a file was not correctly time-aligned, diarisation quality and incorrect prompt intervals were not checked. The Praat interface automatically flagged missing- or duplicated prompts.

3.2.2 Distributing data across multiple sites

A webtool was created to distribute audio recordings and automatically–generated textgrids amongst the annotation team, to gather manually corrected textgrids, and to regularly check consistency across annotators (Szalay et al., 2022a). The front-

end of the webtool offered two interfaces: one for annotators and another for their supervisor. Users were directed to their respective interface based on their role and authentication credentials. The annotator interface allowed downloading of a newly allocated audio file along with the corresponding textgrid and uploading corrected annotation files.

To ensure consistency in corrections, benchmark files with ground truth annotations created by expert phoneticians were inserted at regular intervals (Sect. 3.4). Annotators were blinded as to which files were ground truth files. Once corrections to benchmark files were uploaded, they were automatically scored against the ground truth annotations. An evaluation score was calculated, and a detailed report of the annotation accuracy was generated. This report was then sent to the supervisor via email, along with the score and additional information. (Sect. 3.4). The annotator had to achieve a passing score before proceeding to the next file.

If a passing score was not achieved, the annotator's access to the webtool was temporarily suspended until the supervisor manually reinstated it through the supervisor's interface. All corrected files since the last successful checkpoint were checked manually by the supervisor and corrected if required. Feedback and additional training was provided to support improvement.

The backend of the webtool featured a database that maintained file allocation statistics, annotator scores, and account information. The backend logic was implemented in PHP scripting language with the MySQL database while the front end was developed with HTML5 and JavaScript. The webtool was hosted on a university web server, facilitating access for annotators from diverse locations.

3.3 Manual correction in Praat

Annotators downloaded the pre-screened audio files from the webtool and manually corrected transcription and time-alignment for task-related child speech. A Praat tool was designed to enable efficient manual correction of Task(1) (Boersma and Weenink, 2023; Szalay et al., 2022a). The Praat interface provided step-by-step instructions for manual correction and automated low-level and time-consuming tasks (e.g., opening and saving textgrids), loading the intervals of the sound file likely to contain a target word and skipping the rest. The tool was custom-built to suit the requirements of correcting the single-word production in Task(1) (Szalay et al., 2022a). Annotators received an instruction manual with the annotation guidelines, as well as on-line training in the use of the webtool and the Praat tool. Annotators were required to complete two files with known ground truth annotation prepared by experts before they were allowed to work on new files (Sect. 3.4).

3.3.1 Focus on intervals of interest

During manual correction, the Praat tool automatically loaded intervals of interest – intervals most likely to contain target words – by comparing the prompt tier to the automatic transcription in an iterative process and presented them for the annotators to correct (Fig. 5). In the first iteration, the Praat tool loaded and presented for manual correction the intervals in which the automatic transcription matched the prompt, as

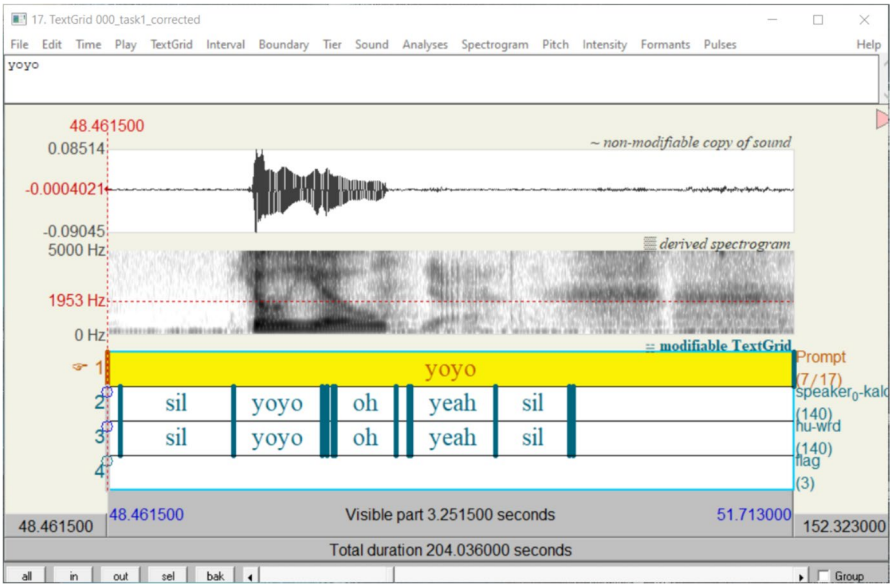

Fig. 5 An example of a target interval prior to manual correction. Tier 1: expected target shown on the prompt tier. Tier 2: automatically generated tier with boundaries placed by NeMo and transcription generated by UNSW ASR. Tier 3: transcription to be corrected by the annotator. Tier 4: markers for noisy targets and targets with mispronunciations to be placed by the annotator

these were the most likely to contain a target word. For instance, the interval labelled with the target word *key* was loaded only if the corresponding prompt interval was also labelled with *key*. When more than one interval label matched the prompt, they were all loaded, presented, and corrected consecutively. The rest of the intervals, i.e., intervals that were transcribed as non-target words or target words that did not match the prompt were skipped. The tool tracked which prompts were found and corrected in the first iteration.

In the second iteration, the script loaded all the word-level intervals that mapped onto prompts not found in the first iteration. For instance, if the target *key* was found in the first iteration, then all other word-level intervals that mapped onto the prompt *key* were assumed to be non-target and skipped. If, however, the target *key* was not found in the first iteration, all intervals identified as the child's speech produced during the prompt *key* were loaded, irrespective of their transcription. If *key* was produced twice, and transcribed as *he* and *Kay* respectively, both were loaded and corrected in the second iteration.

In the third iteration, the script loaded an entire prompt interval if the prompt was not found in either of the previous iterations. The annotator was asked to listen to the entire interval during which a prompt was displayed to identify and add the target word. Multiple target intervals could be added. Target words that were not found in any of the iterations were assumed not to have been produced or were only produced when not prompted, and therefore were not transcribed.

Repeated productions were identified when they were corrected in the same iteration. Repetitions were skipped when they would be corrected in different iterations.

For example, when the target *key* was produced twice, with one instance having correct automatic transcription and the other incorrect, the interval with the correct automatic transcription was corrected in the first iteration, and the one with the incorrect transcription was skipped in the second iteration. This method may have led to the loss of repeated items; however, this was considered an acceptable loss to save annotation time as the task was not designed for repeated target word elicitation.

3.3.2 Steps of manual correction

Each interval was corrected in four steps: interval evaluation, label and boundary evaluation, noise evaluation, and phoneme-level discrepancy evaluation. When the Praat tool first presented an interval to the annotators, the annotators had the options to Accept, Delete (Not Child), or Delete (Not Target) the interval (Fig. 6). The annotators were instructed to only accept an interval if the interval (1) contained the child's voice, and (2) the child's speech matched the prompt. Annotators were instructed to accept targets embedded in continuous speech (e.g., *treasure* in *treasure chest*) and targets produced with errors, as long as they unambiguously matched the prompt (e.g., *three* pronounced as [fwi:]. Three targets (*dinosaur, hippopotamus, rhinoceros*) had acceptable short forms (*dino, hippo, rhino*) and five had acceptable diminutives (*bird - birdy, duck - ducky, fish - fishy, frog - froggy, pig - piggy*). When annotators selected Delete (Not Child) or Delete (Not Target), the Praat tool automatically removed the interval.

After the annotators accepted an interval, they were instructed to either accept or edit the interval's label (i.e., the automatic orthographic transcription) and/or its

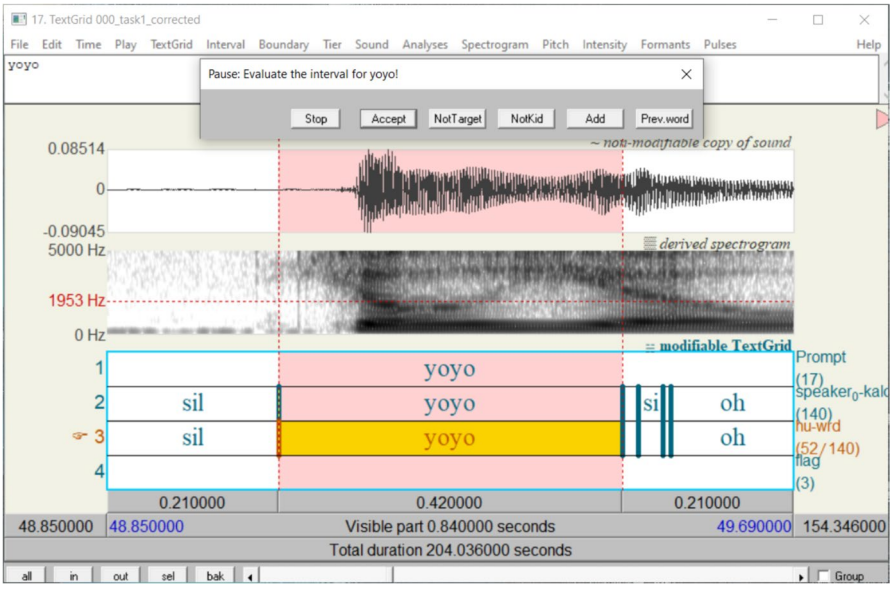

Fig. 6 Step 1 of hand-correction: evaluating the interval. The child's speech shown on the spectrogram matches the expected target shown on the prompt tier (Tier 1), thus the interval is to be accepted

boundaries (Fig. 7). Labels were accepted as correct when they matched both the child's speech and the prompt. Boundaries were considered correct when the interval included the entire target word without any noise or non-target speech and a 50 ms final trailing space for target words ending in stop consonants. For other targets, intervals were accepted irrespective of the length of the trailing space, e.g., a short, 1 ms long trailing space was not lengthened and a long, 250 ms long space was not shortened to save annotation time.

When a label or a boundary was incorrect, annotators were instructed to select edit (Fig. 7). When correcting labels, annotators were instructed to transcribe the child's speech using standard English spelling and grammar (Fig. 8). For example, the target word *rhinoceros* was to be spelled with the letter "r", regardless of whether it was produced with the rhotic or an approximant, such as [w] or [ʊ]; *two eggs* was spelled with the plural marker -s, even when the child produced *two egg*. Verbatim transcriptions were not created.

When correcting boundaries, annotators were instructed to include the entire word produced by the child and no other speech or noise. In particular, they were cautioned to be careful to include final stop bursts which may be cut off by automatic annotation tools due to children producing a longer closure phase in stops than adults (Millasseau et al., 2021). Annotators were instructed to lengthen target intervals when the automatically placed boundary cut off the start or the end of the word and to shorten target intervals when the automatically placed target boundaries included noise (e.g., other speaker, tapping, coughing) (Fig. 8). When a boundary was edited, annotators

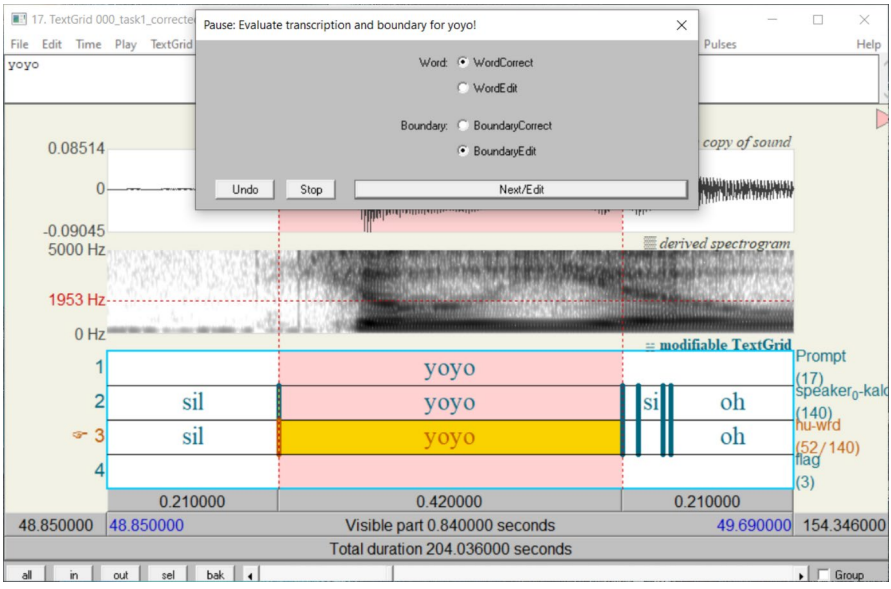

Fig. 7 Step 2 of hand-correction: evaluating the label and the boundaries. The automatic transcription on Tier 2 matches the child's speech shown on the spectrogram, thus it is to be accepted. The automatically placed boundary on Tier 2 excludes the last vowel of *yoyo* shown on the spectrogram, thus the boundaries are to be edited

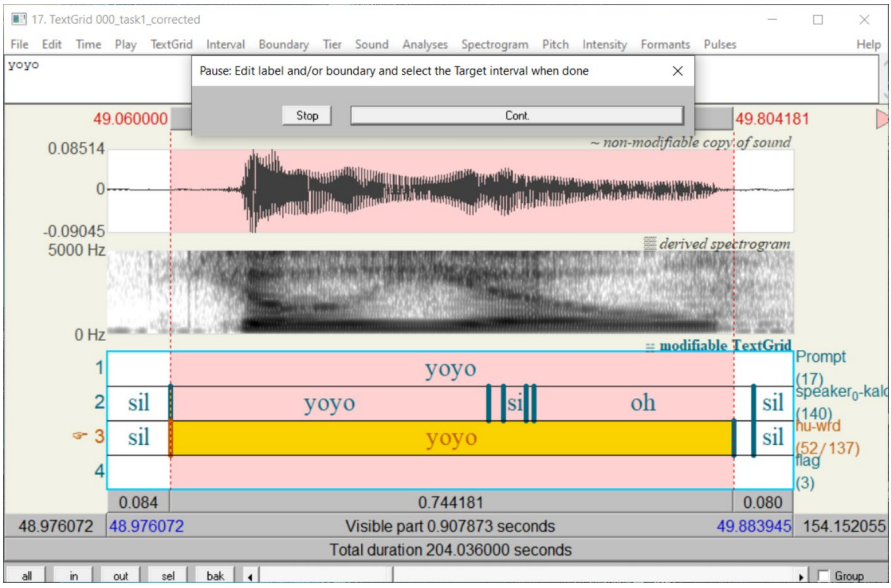

Fig. 8 Step 3 of hand-correction: editing the interval. The annotator has edited the final boundary of *yoyo* on Tier 3 to include the entire word

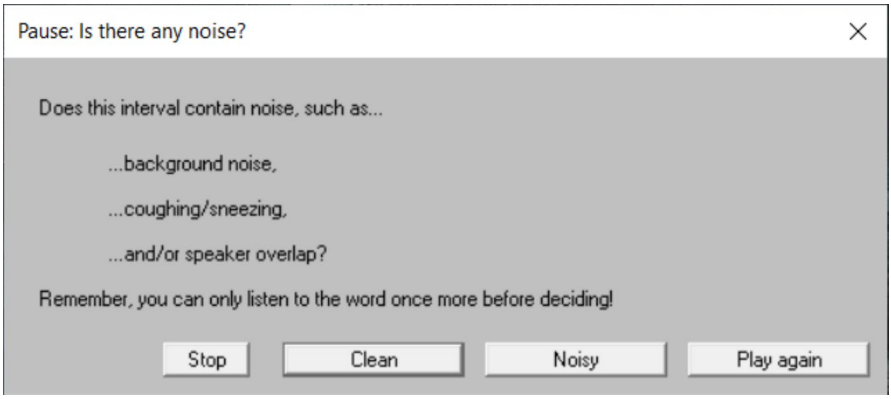

Fig. 9 Step 4 of hand-correction: flagging noisy tokens. The annotator is prompted to evaluate whether the target contains noise, such as speaker-overlap, coughing, or tapping on the microphone

were instructed to leave a 20 ms trailing space at the start and end of the word unless the word ends in a stop consonant, in which case 50 ms trailing space was required.

Having completed the necessary corrections, annotators were asked to decide whether the interval contained any noise (e.g., overlap between the interviewer and the child, background noise, Fig. 9) and whether the token contained any phoneme-level insertions (e.g., *skirts* for *skirt*), deletions (e.g., *four egg* instead of *four eggs*), or substitutions (e.g., *tooth* produced with a final /f/ instead of /θ/, Fig. 10). Annotators were instructed to flag any phoneme-level discrepancy from adult production irre-

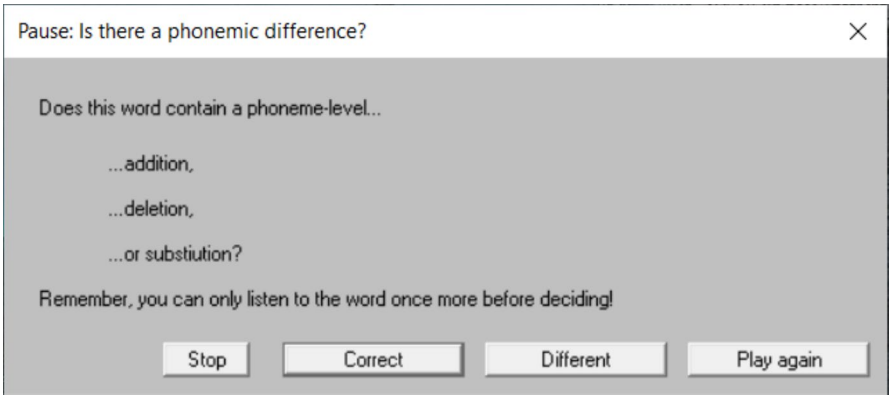

Fig. 10 Step 5 of hand-correction: flagging phoneme-level mispronunciations. The annotator is prompted to evaluate whether the target contains a phoneme-level deletion, addition, or substitution

spective of it being age-appropriate (e.g., substituting [w] for /ɹ/ was flagged at each instance, regardless of age). Evaluating the presence of noise and phoneme-level discrepancies was a binary (Yes/No) decision. To prevent the annotators from spending too much time on noise and discrepancy evaluations, the Praat tool only allowed the target word to be played twice for noise and twice for discrepancy evaluation (four times in total) and automatically closed the sound and the textgrid while waiting for the noise and discrepancy decisions. The Praat tool automatically loaded the next interval once all four steps for a given interval were completed.

To improve annotation quality, the Praat tool verified that the annotator followed the instructions as closely as possible. A notification was triggered when the interval label did not match the prompt and/or when the annotator's evaluation did not match what they have done. For example, when the annotators evaluated the automatically generated label as correct yet they edited it, they received a pop-up notification to correct their error(s). Erroneous edits were not saved and the annotator was not allowed to move onto the next interval until all checks were passed. When all checks were passed, the Praat tool saved the final edits and marked the interval as corrected, allowing annotators to exit Praat any time. At the next launch, the script loaded an unfinished sound file - textgrid pair at the first uncorrected interval.

When the annotator corrected the textgrid for the entire sound file, the Praat tool automatically created a clean and corrected textgrid free from unnecessary annotations by removing all child labels in which the label did not match the prompt. To create an easily searchable output, the Praat tool generated a csv file, listing the label, the start- and the end time of every on-prompt target. Annotators manually uploaded the clean and corrected textgrid, together with the csv file to the webtool.

3.4 Reliability checks

Inter-rater reliability is typically evaluated by all annotators correcting the same set of speakers (customarily 20% of the data). Due to the large number of speakers, a

20% cross-correction was not possible, as it would have required 124 speakers being marked by all annotators.

Therefore, a ground-truth approach was chosen to achieve consistency. Six ground truth sound files were identified consisting of four older (10-12 years, M = 2, F = 2) and two younger children (3-5 years, M = 1, F = 1). Four of the children were typically developing, and two had current speech sound disorders (one older male and one older female). An expert phonetician and a speech pathologist (Authors Tünde Szalay and Kirrie Ballard) manually corrected the benchmark files independently from each other and compared their correction. In cases of disagreement, consensus was reached through discussion.

Ground truth files were inserted automatically by the webtool after every 10th file (Sect. 3.2). The same ground truth files were re-used with dummy participant numbers. The number of unique targets identified by the annotator was compared to the number of targets identified in the ground truth correction. The number of identified targets tested whether the annotator found all the target words in the audio data. For every matching target, overlap rate between the annotator's work and the ground truth was calculated. Overlap rate was calculated using the time shared between ground truth and current annotation (*Dur Shared*), the duration of the ground truth annotation (*Dur GT*), and the duration of the current annotation (*Dur Current*, Eq. 1, (Gonzalez et al., 2020; Paulo and Oliveira, 2004)). Overlap rate ranged from 1 (complete agreement, 100% overlap) to 0 (no agreement, 0% overlap) and penalised too long and too short intervals equally.

$$Overlap = \frac{DurShared}{DurGT + DurCurrent - DurShared} \tag{1}$$

Passing rate for annotators was calculated automatically from the number of target words identified and from the overlap rate (Sect. 3.2). A report was generated automatically and emailed to the supervisor (Fig. 11). To achieve a passing score, annotators had to find all but one target word and have less than than 5% of targets with an overlap rate below 0.85. If a passing rate was achieved, annotators were automatically and immediately allowed to access their next file without being notified that they have completed a reliability check.

If a passing rate was not achieved, the ground truth file was reviewed. If the fail was due to missed targets, all the corrected files after the last successful checkpoint were reviewed (Fig. 1).

When the fail was due to low overlap rate, targets with overlap rate below 0.85 were evaluated on a three-point scale: Phonetic Errors, Instruction Errors, and Opinion. When the annotator's interval only contained part of the target word, e.g., missed a coda burst, or contained non-target speech, the error was evaluated as Phonetic Error. When the annotator's interval contained the entire target word and only the target word, but was not compliant with the instructions, the error was labelled as Instruction Error.

Lastly, when a low interval rate was caused by ambiguity in the acoustic signal, the error was rated as Opinion. For example, target word onset was considered ambiguous, when the target was preceded by a sound ambiguous between non-target

```
<Annotator Name>

Detailed Report
---------------

Found Target Ratio = 99.0%
Total Number of Targets = 137
Number of Found Targets = 136
Number of Missed Targets = 1
Number of Extra Targets = 0


------------------------
Detail of Missed Targets
------------------------

Label | Start-Time | End-Time
------+------------+---------
spade | 679.85     | 680.41

Overlap_ratio_mean = 0.99
Overlap_ratio_median = 1.0
Overlap_ratio_max = 1.0
Overlap_ratio_min = 0.73

Percentage of intervals with Overlap ratio < 0.85 = 1.0%
Number of intervals with Overlap ratio < 0.85 =  2


-----------------------------------------
Detail of intervals of Overlap ratio < 0.85
-----------------------------------------


Label   | GT-Start-Time | GT-End-Time | AN-Start-Time | AN-End-Time | OR
--------+---------------+-------------+---------------+-------------+-----
yoyo    | 311.04        | 311.81      | 311.15        | 311.95      | 0.73
coconut | 614.60        | 615.36      | 614.60        | 615.23      | 0.83
```

Fig. 11 Sample of the automatically generated report for a passing ground truth annotation. As all but one of target words were found and less than 5% of targets had an overlap rate below 0.85, the annotator passed reliability check

noise, false start, or pre-voicing. Vowel-final target words often ended in a sound that was ambiguous between a de-voiced vowel or an outbreath, making target offset challenging to find. In ambiguous tokens, the boundary in question was compared to the acoustic signal and the instructions. If the boundary was placed in line with a plausible acoustic landmark and the instructions, the target was rated as Opinion, indicating that the annotator relied on an acoustic landmark that was correct but different from the ground truth (Fig. 12).

After rating the errors, the percentage of targets with an overlap rate below 0.85 was re-calculated using only Phonetic and Instruction Errors. If the annotators reached passing rate, they were given feedback on their errors and asked to continue annotation. If the proportion of targets with a low overlap rate still exceeded 5%, all files between the last passed ground truth were reviewed and annotators were only allowed to continue after successfully completing a second ground truth file.

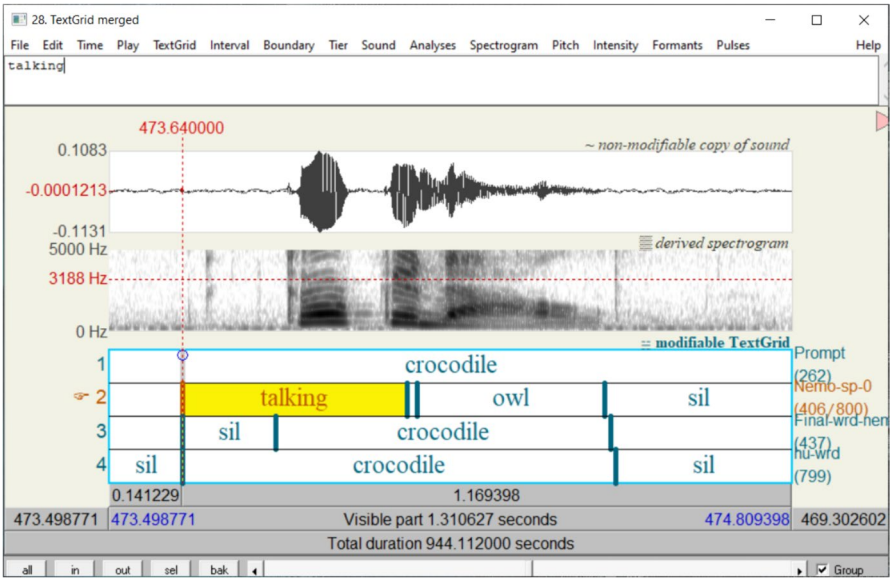

Fig. 12 Annotator's correction (Tier 4) compared to ground truth correction (Tier 3) for the prompt *crocodile* (Tier 1) together with automatic transcription (Tier 2) for a difference in opinion. The spike preceding the target word is ambiguous between a false start and noise

4 Workflow evaluation

4.1 Outcomes of automatic textgrid generation

Pre-processing was attempted for all 620 Task(1) files and successfully returned a textgrid for 454 (73%, Fig. 13). Automatic pre-processing failed to generate textgrids when the speech data were recorded using a paper protocol for prompt presentation instead of the tablet, the audio file was named incorrectly during data collection, or when time-alignment was not possible due to failed tone detection or missing timestamps. Automatically processed files were pre-screened, showing that 394/454 (87.0%) textgrids were of good quality. Six files were considered to be too noisy, 28 were aligned incorrectly, and 26 were diarised poorly. Nine recordings that failed pre-screening had been hand-corrected prior to the introduction of the pre-screening step, suggesting these recordings can be hand-corrected, albeit possibly more slowly and with more effort by the annotating team. Methods for salvaging excluded files are currently being explored. For example, when recordings from the primary microphone were evaluated as too noisy, audio files recorded via the far field microphone and the first and second directional microphones are being evaluated for substitution.

In the files passing pre-screening, on average 2.9 (minimum = 2, maximum = 5) speakers were identified. The expected number of speakers in the audio was 3 (child, interviewer, model speaker). The lower than expected minimum suggests than some speakers were missed by NeMo, potentially due to not producing enough speech. The higher than expected maximum is consistent with more than expected number of speakers being present during recording (e.g., caretakers, siblings) and with NeMo

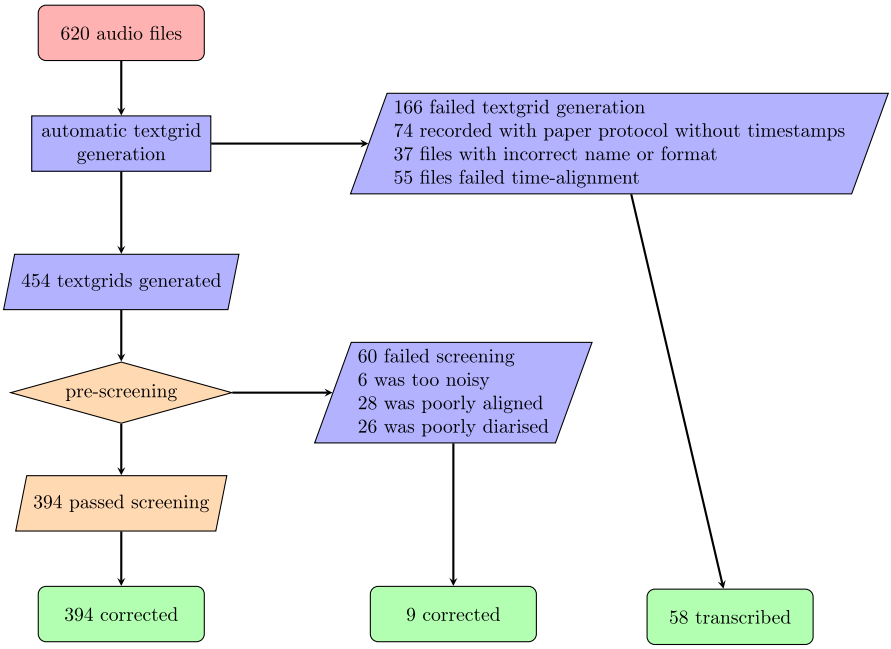

Fig. 13 Automatic pre-processing outcomes. Red: start point. Green: end point for corrected or transcribed files. Blue: automated process Yellow: manual process. Diamond: decision point. Trapezium: interim outcome

Table 2 Number of orthographic transcriptions compared to the total number of children in the AusKidTalk by speaker age and gender, including the files transcribed without ASR-assistance

Age (years)	Girls		Boys		Total	
3–5	81/103	(79%)	69/100	(69%)	150/203	(74%)
6–8	76/103	(74%)	90/114	(79%)	166/217	(77%)
9–12	75/93	(81%)	70/107	(65%)	145/200	(73%)
Total	232/299	(78%)	229/321	(71%)	**461/620**	**(74%)**

Percentage transcribed in parenthesis

erroneously identifying one speaker's voice as belonging to two separate speakers, e.g., due to varying loudness of the speaker. A total 170 prompt intervals were missed in the 394 files passing pre-screening, yielding an average of 0.43 missing prompts per file. A total of 151 prompts were duplicated, yielding an average of 0.38 duplicated prompts per file. Annotators successfully added all missing prompts and removed all duplicated prompts.

A total of 461 files were orthographically transcribed, either by hand-correcting automatic transcription (403 files) or without automatic transcription, from scratch (58 files, Fig. 13). The number of transcribed files ranges from 65%-81% between age groups and gender, with no apparent difference between age groups and genders (Table 2). To further examine the effect of age and gender on the AusKidTalk workflow's performance, a generalised linear model was constructed with the binomial family in R (R Core Team, 2025). The response variable was Tran-

scription Completion (binomial, transcribed coded as 1 and not transcribed as 0); the fixed effects were Age in years (numeric) and Gender (boys compared to girls as a baseline) with an interaction term for Age × Gender; random effect for children was not included as each child contributed one Task(1) recording. None of the fixed effects reached the threshold of significance, showing no effect of age or gender on transcription completion (Age: $\beta = 0.048, z_{0.05} = 0.878, p = 0.3801$; Gender: $\beta = 0.078, z_{0.55} = -0.142, p = 0.8872$; Age × Gender: $\beta = -0.043, z_{0.07} = -0.598, p = 0.5497$). That is, there was no evidence for the workflow performing better on older children compared to younger children or on the speech of boys relative to girls.

4.2 Diarisation and ASR quality

The quality of NeMo Diarisation and UNSW ASR transcription was evaluated by comparing automatically generated transcriptions to their manual corrections on a sample of 380 children, including children with autism spectrum disorder and/or speech sound disorder. In a sample of 380 files, annotators evaluated 85,436 intervals and identified 61,371 target words (Fig. 14). Out of the total of 85,436 intervals, 6.2% were rated as "Not Child" and deleted when the NeMo diarisation tool incorrectly mapped the interval onto the child's tier. Adult speakers may have been mapped onto the child's tier by NeMo due to a false positive in child identification. Alternatively, non-child words may have been added to the child's tier by UNSW ASR, when it automatically added all recognised words to all speaker-tiers when a prompt interval did not overlap with any speaker-intervals. 4.5% of the total intervals were added

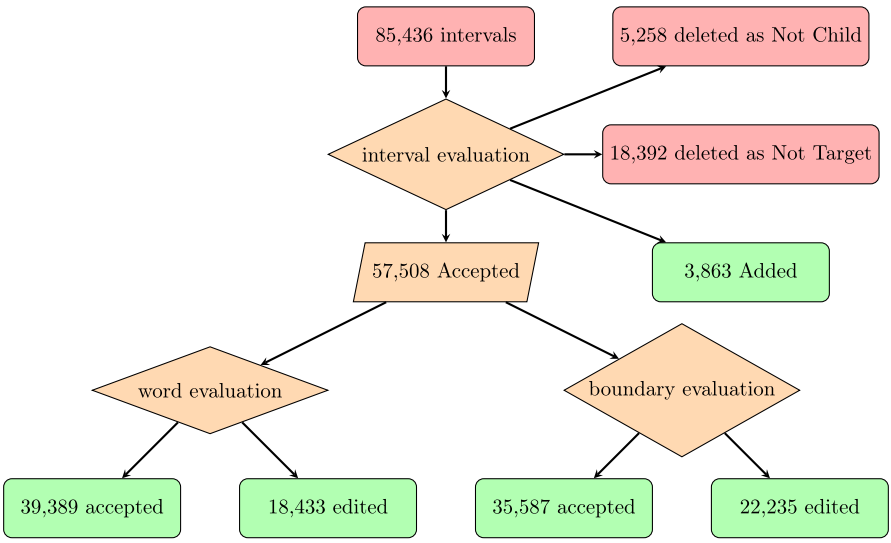

Fig. 14 Outcomes of manually evaluating the automatically identified intervals. Red: start point and end point of deleted intervals. Green: end point. Yellow: manual process. Diamond: decision point. Trapezium: interim outcome

Table 3 Mean WER (%) by Age Group, and Gender, calculated using "Accepted" intervals and excluding compounds and counting tasks

Age (years)	Girls	Boys	Total
3–5	32.9	29.7	31.6
6–8	16.1	15.7	15.9
9–12	12.3	12.5	12.4
Total	20.6	18.6	**19.7**

Number in bold shows total WER

Table 4 Mean overlap by Age Group, and Gender, calculated using "Accepted" intervals and excluding compounds and counting tasks

Age (years)	Girls	Boys	Total
3–5	0.87	0.88	0.87
6–8	0.94	0.95	0.95
9–12	0.96	0.97	0.96
Total	0.92	0.94	**0.93**

Number in bold shows total overlap

manually when the NeMo diarisation did not identify a target word as the child's speech, suggesting a 4.5% false negative for diarisation (Fig. 14).

When an automatically generated interval was accepted as a target word produced by the child, the automatically generated interval was evaluated for transcription- and alignment accuracy. Word error rate (WER) and overlap rate were calculated using accepted target words by comparing the automatically generated interval to its manual correction (Sect. 3.3). That is, a total of 57,508 target intervals were used to calculate WER and overlap rate (Fig. 14). Out of the total of 57,508 target words, 68% had accurate automatic transcription, and 32% had transcription errors, yielding a WER of 32%. Mean overlap rate was 0.86.

WER increased and overlap rate was reduced as two compound words (*spider-web*, *bellybutton*) and the counting phrases (e.g., *one egg, four o'clock*) were always mapped onto two separate intervals by the automatic transcription and onto a single interval during hand-correction (*spider* and *web*; *belly* and *button*, *one* and *egg*; *four* and *o'clock*). Transcriptions and boundaries of these compounds and the phrases were always evaluated as incorrect and had to be edited to include the entire compound or the phrase. When the two compounds and all counting phrases were excluded prior to calculating WER and overlap, WER reduced to 19.7% and overlap rate increased to 0.93 (Tables 3 and 4). The UNSW WER was expected to be high, as UNSW ASR was trained on accent-mismatched child speech (Sect. 3.1.3). The current 19.7% WER is comparable to that of other child ASR tools developed on adult speech, highlighting the need to use accent-matched child data to train child ASR (Bhardwaj et al., 2021; Kathania et al., 2021, 2022; Shivakumar and Georgiou, 2020; Yadav et al., 2018).

The effect of age and gender on WER and overlap were further examined by constructing two generalised linear mixed models in R (R Core Team, 2025). For WER, the response variable was Word Error (binomial, correct automatic transcription coded as 0 and erroneous automatic transcription coded as 1), with the family binomial in the *lme4* library; $p - values$ were calculated using Satterthwaite's degrees of

Table 5 Audio- and annotation time for 380 hand-corrected and 58 transcribed files

	Annotation method	Audio duration	Play time	Annotation time
Total (hours)	Manual correction	136.6	18.5	Unknown
Average (minutes)		22	3	90
Total (hours)	Transcription	35.4	35.4	223
Average (minutes)		36	36	230

freedom method in the *lmerTest* library (Bates et al., 2015; Kuznetsova et al., 2017). For Overlap, the response variable was Overlap (beta-transformed to include 0 and 1 using a weighted average method) and the family beta was used in the *glmmTMB* library (Brooks et al., 2017; Smithson and Verkuilen, 2006). In both models, the fixed effects were Age in years (numeric) and Gender (comparing boys to girls as a baseline) with an interaction term for Age × Gender; a random effect for child was added.

Age had a significant negative effect on WER, consistent with decreased WER and better ASR performance on older children's speech relative to younger children (Age: $\beta = -0.212, z_{0.01} = -15.718, p < 0.001$). Gender and the Age × Gender interaction did not reach the threshold of significance (Gender: $\beta = -0.24, z_{0.15} = -1.58, p = 0.114$; Age × Gender: $\beta = 0.025, z_{0.02} = 1.251, p = 0.211$). Age had a significant positive effect on Overlap, consistent with increased Overlap and better diarisation performance on older children's speech compared to younger children (Age: $\beta = -0.058, z_{0.004} = 12.51, p < 0.001$). Gender and the Age × Gender interaction did not reach the threshold of significance (Gender: $\beta = -0.092, z_{0.05} = 1.72, p = 0.0856$; Age × Gender: $\beta = -0.009, z_{0.01} = -1.28, p = 0.1990$)

4.3 Reducing the time-need of manual speech transcription

The Praat tool's efficiency in reducing hand-correction time and cost was evaluated on a sample of 380 children – including children with autism spectrum disorder and/or speech sound disorder – with an average Task(1) length of 22 min per file and a total duration 136.6 h (Table 5). The total audio duration of 136.6 h was automatically reduced to 18.5 h of play time, with an average play time of 3 min per file, as annotators only played the parts of the audio loaded by the Praat tool. Annotators reported correcting one file in 90 min on average.

Two experienced annotators (Author Renata Huang and a speech-language pathologist) transcribed 58 files recorded using a paper protocol for prompt presentation instead of the tablet. For these files, the experienced annotators listened to the entire recording consisting of five speech tasks and identified and extracted the tasks manually using Praat textgrids. Identifying five tasks took 5 min on average. Having identified the tasks, the annotators loaded Task(1) audio to Praat, and listened to the recording to transcribe every target word produced by the child onto a Praat textgrid. Lastly, the annotator marked noisy and mispronounced targets. Transcription was aided by a Praat script that reminded the annotator the order in which prompts were presented and provided an interface for flagging noisy and mispronounced target.

Transcribing a file with no automatically generated transcription took 230 min (3.8 h) on average, considerably longer than manual correction of automatically generated transcriptions (90 min).

4.4 Reliability of manual correction

Annotators completed 49 ground truth file checks, achieving a passing score on 41 files (83.6%). The 41 consistent ground truth submissions, on average, identified 99.9% of all targets with 2.3% of tokens with an overlap rate below 0.85. The eight inconsistent ground truth submissions failed due to the low overlap rate, never due to missed targets. Therefore, alignment errors in the eight failed ground truth submissions were re-scored manually. After re-scoring, two files were rated as consistent, yielding a total of 43/49 (87.7%) consistent ground truth files.

In the six failed files, typical phonetic alignment errors were missed coda bursts, difficulties identifying the start and end of fricatives, and identifying target words in connected speech (e.g., the target *spade* when the child produced *spade'n'garden*). As phonetic errors may reduce data quality for ASR training, the annotator's work between the prior successful reliability check and the failed one was revised and the annotator was given an additional ground truth file for re-training.

Typical instruction errors were caused by annotators shortening automatically placed intervals despite the intervals not containing noise or non-target speech. Although instruction errors are not expected to adversely impact data quality, they do increase annotation time and cost, therefore annotators were reminded to follow the annotation instructions and were given an additional ground truth file for training.

5 Efficiency, generalisability, and lessons learnt

When building the AusKidTalk corpus, we aimed to remedy the lack of AusE child speech corpus and to solve the circular problem of needing ASR tools trained on AusE-speaking children for annotating AusKidTalk. In this paper, we demonstrated that a reliable pipeline for semi-automating transcription can be built through designing data collection methods that consider how speech technology can be leveraged for transcription. The corpus was co-designed by engineers developing data collection methods with high-frequency tones facilitating task- and prompt identification to reduce annotator burden, and by phoneticians and speech-language pathologists designing speech tasks suitable for children. The resulting corpus contains speech data from 620 children and covers unique properties of AusE. To transcribe the data, the AusKidTalk annotation workflow concatenated multiple automatic tools in a step-by-step manner and augmented it with hand-correction, resulting in hand-correcting speech data from 403 children (Figs. 1 and 13). In addition, 58 children's speech was transcribed without being assisted by ASR tools for comparison.

5.1 Workflow benefits

The main benefit of automatic pre-processing was more efficient hand-correction time compared to transcribing audio without being asssited by ASR tools. Efficiency was achieved by reducing the length of audio that annotators had to listen to by focusing on key intervals of interest. Intervals of interest were identified in three steps: (1) identifying the interval during which prompts were displayed; (2) automatically identifying and transcribing the child's speech; and (3) comparing the prompt to the child's speech to find a match. These three steps allowed annotators to identify targets by only listening to 13.5% of audio: 18.5 h instead of 136.6 h. The three steps; however, were taken at different points during corpus design: timestamps were recorded when data were collected, speaker identification and automatic transcription took place during automatic analysis, and comparison of prompt and automatic transcription was part of the manual correction procedure. That is, successful completion of initial steps had a strong impact on the efficiency of manual correction, demonstrating that strategic data collection with rigorous logging of stimulus presentation times can assist orthographic transcription when written prompts were not provided, allowing for the inclusion of pre-literate children. The workflow also shows that the time gains of using ASR to assist transcription, demonstrated on adult speech, also apply to child speech despite the limitations of child ASR (Bazillon et al., 2008).

The pre-processing workflow concatenated multiple tools to automatically generate transcription, allowing for updating the workflow with state-of-the-art tools as they became available. For instance, IBM-Watson Diarisation was replaced with NeMo during the initial phase of the project, and UNSW ASR may be replaced by better performing ASR models as they become available. Although three tools were concatenated (high-frequency tone detection, NeMo speaker diarisation, and UNSW ASR), cascade errors were only observed at two levels: at data collection (e.g., file naming) and during high-frequency tone detection leading to failed Task(1) extraction. In contrast, no cascade errors were observed for speaker-diarisation and automatic word recognition: when a task file was successfully extracted, the textgrid was generated with speaker diarisation and automatic transcription.

Manual correction was necessary due to the quality of automatic transcription, but is prone to inconsistencies due to human errors. The Praat workflow streamlined hand-correction to reduce human errors, reliability was monitored, and training was provided consistently throughout hand-correction. As a result, overall good reliability was achieved with annotators achieving reliable results on 43/49 ground truth files. Manual correction was carried out by annotators with a background in linguistics and/or speech-language pathology, highlighting the need for experts in speech science in corpus building.

Ongoing work explores avenues for manual correction of speech data where automatic pre-processing was not possible or performed poorly. The workflow cannot be deployed on the recordings collected using a pencil-and-paper protocol due to tablet failure, therefore, such files are being transcribed without the help of automatically generated transcriptions at considerably higher costs per file. For data excluded due to poor audio quality on the primary microphone, the audio quality on the secondary microphones will be examined. Textgrids excluded due to incorrect time-alignment

will be analysed manually, while data excluded due to poor diarisation may be hand-corrected as-is; however, poor diarisation might increase hand-correction time and costs. Manual data cleaning, such as renaming files and adding timestamps, is time-consuming and labour intensive, but it will increase the number of annotated Task(1) files. Manual data cleaning is also expected to have a roll-on effect for the remaining tasks: when one child's audio files are renamed, then data for all five tasks become analysable using a similar workflow.

5.2 Workflow generalisability

The Task(1) workflow generalised well to Task(3), which elicited narrative speech using video- and picture prompts (Szalay et al., 2025, 2026). The existing pre-processing workflow identified Task(3) audio and generated time-aligned orthographic transcription for 429 Task(3) files automatically, out of which 297 passed pre-screening. Slightly poorer performance of the pre-processing pipeline on Task(3) compared to Task(1) was attributed to the interviewer contributing a limited amount of speech, and thus speaker diarisation performing less well (Szalay et al., 2025, 2026).

The textgrids were distributed among annotators for manual correction using the same webtool. Manual correction required developing a new interface meeting different requirements for correcting a narrative task, leading to additional investment of time and resources (Szalay et al., 2026). In the narrative task, the hand-correction interface was designed to identify the child's turns (i.e., the child's speech, uninterrupted by the interviewer) as opposed to target words, and annotators were prompted to flag hesitations and partial words, typical in continuous speech, as opposed to phoneme level errors (Szalay et al., 2026). Generalisability to Task(3) suggests that the workflow may be beneficial for novel child speech corpora.

To further test generalisability and increase the reuse of the pipeline, the pre-processing and hand-correcting workflow may be adapted for Task(5), the non-word repetition task using audio prompts. In the pre-processing pipeline, high-frequency tone detection is predicted to identify Task(5) with an accuracy comparable to Task(1) and Task(3) identification. NeMo is expected to perform equally well on Task(5) as on Task(1), due to Task(5) containing a comparable amount of speech from the model speaker providing the audio prompt and the child repeating it. UNSW ASR, used for orthographic transcription in the pre-processing pipeline, however, would require the non-words transcription and pronunciations to be added to the lexicon and the hand-correction interface of Task(1) would need to be modified to accommodate the new target words (e.g., *contramponist* [kəntɹæmpənəst]), requiring further time and resources.

5.3 AusKidTalk use cases

AusKidTalk was designed for future development of automatic speech recognition and analysis algorithms and tools, as well as for answering research questions related to speech and language development and disorders (e.g. development of Australia-specific normative data). The accurate orthographic transcription generated in this workflow is a necessary prerequisite for multiple avenues of future work, such as

ASR-training. For instance, preliminary experiments indicate that fine-tuning ASR on the single word production task brings a small improvement to ASR performance on the narrative task. As age groups are represented evenly among the orthographically transcribed Task(1) files, novel ASR can be trained using younger children's data in the future, potentially leading to larger gains in the younger 3–6 age group, where UNSW ASR currently performs less well. ASR trained on AusKidTalk is expected to perform less well on multicultural AusE or on the speech of children who speak English as a second language, as only children who had at least one parent completing all their schooling in Australia were included in AusKidTalk (Ahmed et al., 2021). Future research may scale AusKidTalk using the same stimuli and data collection protocol to augment the corpus with data from a more diverse selection of children or include speech data from multicultural AusE-speaking adolescents for ASR development (Cox and Penney, 2024).

AusKidTalk was used to fine-tune an automatic mispronunciation detection tool, originally developed for detecting mispronunciations in non-native English speech relative to American English, to detect developmental and pathological child speech errors in AusE-speaking children (Hu et al., 2026; Shahin et al., 2025). The tool analyses phoneme productions of children, compares them to the expected lexical forms, and categorises each phoneme produced by the child as "correct", "substituted", "inserted", or "deleted", relative to the expected pronunciation (Ballard et al., 2025; Hu et al., 2026). The mispronunciation detection tool was tested on 12 children from AusKidTalk (typically developing = 6, children with speech sound disorders = 6), by comparing the tool's evaluation to ground truth evaluation provided by expert speech-language pathologists (Hu et al., 2026). The tool reached an agreement rate above 85% on typically developing and disordered child speech; however, agreement varied depending on the specific error-type (substitution, deletion, insertion) (Hu et al., 2026).

In addition, orthographic transcription is necessary prior to phonetic transcription, required for future phonetic and phonological analysis e.g.,, (Arciuli et al., 2019; Buttigieg et al., 2021; Gibson et al., 2023). Therefore, orthographic transcription of the single word production task will enable research using the large AusKidTalk dataset for multiple research projects, ranging from describing patterns of language acquisition, characteristics of child speech in children with autism spectrum disorder and/or speech sound disorder, and will provide a snapshot of the acoustic-phonetic characteristics of AusE-speaking children.

5.4 AusKidTalk data and workflow availability statement

Speech corpora provide crucial resources in phonetics and speech technology through their reuse value for novel research (Liberman, 2019). Therefore, speech data and transcription are available for research purposes only under a data-custodian model upon request to the last author. Speech data and transcription are not available publicly to protect participants' data and privacy, because, due to advancements in automatic speaker recognition and identification, speech can be used to identify an individual, and is thus classified as personal information and data (ARDC, 2022; Jahangir et al., 2021; OAIC, 2026). Code created for the automatic pre-processing

tools and manual correction, together with citation information are available via the authors' GitHub.[2]

6 Conclusion

In this paper, we described building AusKidTalk – the first AusE child corpus suitable for developing automatic speech recognition tools – and provided a workflow for efficient and reliable orthographic transcription of target words time-aligned with children's speech. We collected data from 620 children, automatically generated textgrids for 454, and hand-corrected or transcribed 461. Our workflow saved a considerable amount of time: a sample of 380 children resulted in a total duration of 136.6 h of speech data, but annotators only listened to 18.5 h of speech to identify 61,371 target words. This reduced hand-correction time to approximately 90 min per child compared to the 4 h required for transcription unassisted by ASR. Hand-corrections were overall of good quality, as annotators' transcriptions consistently matched the ground truth. The annotated data are currently being used to fine-tune ASR tools for AusE-speaking children, thus improving automatic transcription for the remaining tasks of AusKidTalk.

Our workflow may generalise to other child corpora presenting similar problems, as we showed that NeMo diarisation can separate the child speech from that of the adults, and ASR-tools can expedite transcription. Our workflow is particularly useful for single-word production tasks, which is one of the most commonly used tasks in child data collection, given that the availability and limitations of speech technology tools are considered when designing the data collection protocol. The workflow can be modified for different target words and extended to include non-words. The workflow code is available via the authors' Github and data from the last author upon request under a data custodian model.

Acknowledgements 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 Committee, Approval Number HC190320.

Author contributions Author TS prepared the initial manuscript, conducted data analysis and interpretation, and contributed to the development of new software tools. MS and ST contributed substantially to the development of new software tools and revised the manuscript critically for important intellectual content. ZN contributed to the development of new software tools. RH contributed substantially to data acquisition. JA, EB, FC, KB, and BA. made substantial contributions to the conceptualisation, and design of the work, revised the manuscript critically, and acquired funding. All authors have reviewed, contributed to, and approved of the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data availability No datasets were generated or analysed during the current study.

[2] Pre-processing tools: https://github.com/mostafashahin/AusKidTalkV3 Pre-screening and hand-correcting tools: https://tuendeszalay.github.io/scripts/AusKidTalk/akt_overview

Declarations

Conflict of interest Co-authors Joanne Arciuli, Elise Baker, Felicity Cox and Kirrie Ballard own the intellectual property of the Speech Assessment for Australia Children (STAC) that is mentioned in the manuscript. They have no financial interests. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmed, B., Ballard, K.J., Burnham, D., Tharmakulasingam, S., Mehmood, H., Estival, D. & Ambikairajah, E. (2021). AusKidTalk: An auditory-visual corpus of 3-to 12-year-old Australian children's speech. *Proc. of INTERSPEECH* (pp. 3680–3684). https://doi.org/10.21437/Interspeech.2021-2000

Arciuli, J., & Ballard, K. J. (2017). Still not adult-like: Lexical stress contrastivity in word productions of eight-to eleven-year-olds. *Journal of Child Language, 44*(5), 1274–1288. https://doi.org/10.1017/S0305000916000489

Arciuli, J., Ballard, K. J., Phillips, K., & Vogel, A. (2019). Using MAUS to investigate children's production of lexical stress. *Proc. of the 19th International Congress of Phonetic Sciences* (pp. 2470–2474)

ARDC. (2022). Publishing sensitive data. *Zenodo*. https://doi.org/10.5281/zenodo.7259742

Baker, E., Cox, F., Arciuli, J., & Ballard, K. J. (2019). Speech test for Australian children (STAC)

Ballard, K. J., Djaja, D., Arciuli, J., James, D. G., & van Doorn, J. (2012). Developmental trajectory for production of prosody: Lexical stress contrastivity in children ages 3 to 7 years and in adults. *Journal of Speech, Language, and Hearing Research, 55*(6), 1822–1835.

Ballard, K. J., Robin, D. A., McCabe, P., & McDonald, J. (2010). A treatment for dysprosody in childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research, 53*(5), 1227–1245.

Ballard, K. J., Shahin, M., & Ahmed, B. (2025). AI-driven speech therapy tool for assisting clinicians in diagnosis and therapy. *Digital Health Week*, *4*(1).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M. & Wong, M. (2005). The PF_STAR children's speech corpus. *Proc. of the* 9^{th} *European Conference on Speech Communication and Technology* (pp. 2761–2764). Lisbon: ISCA.

Bazillon, T., Esteve, Y., & Luzzati, D. (2008). Manual vs assisted transcription of prepared and spontaneous speech. In *Proc. of International Conference on the Sixth Language Resources and Evaluation* (pp. 1067–1071).

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*, 85–100.

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

Bhardwaj, V., Kukreja, V., & Singh, A. (2021). Usage of prosody modification and acoustic adaptation for robust automatic speech recognition (ASR) system. *Revue d'Intelligence Artificielle, 35*(3), 235–242.

Boersma, P., & Weenink, D. (2023). *Praat 6.3.17.* http://www.fon.hum.uva.nl/praat/

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*(2), 378–400.

Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R. & Hajek, J. (2011). Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box. *Proc. of INTERSPEECH* (pp. 848-851).

Buttigieg, L., Grech, H., Fabri, S. G., Attard, J., & Farrugia, P. (2021). Automatic speech recognition in the assessment of child speech. In *Manual of Clinical Phonetics* (pp. 508–515). Milton Park: Routledge.

Chen, N. F., Tong, R., Wee, D., Lee, P. X., Ma, B., & Li, H. (2016). SingaKids-Mandarin: Speech corpus of Singaporean children speaking Mandarin Chinese. In *Proc. INTERSPEECH* (pp. 1545–1549).

Cole, R., & Pellom, B. (2006). University of Colorado read and summarized story corpus. University of Colorado. TR-CSLR-2006-03 Technical report.

Cox, F., & Penney, J. (2024). Multicultural Australian English-the new voice of Sydney. *Australian Journal of Linguistics, 44*(2–3), 200–219.

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

Gao, J., Li, A. & Xiong, Z. (2012). Mandarin multimedia child speech corpus: Cass_Child. In *Proc. of International Conference on Speech Database and Assessments* (pp. 7–12).

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

Gibson, A., Cox, F. & Penney, J. (2023). Acquiring allophony: GOOSE and SCHOOL vowels in the speech of Australian children. In *Proc. of International Congress of Phonetic Sciences* (pp. 3750–3754).

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard, 6*(1), 20190058.

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

Hämäläinen, A., Cho, H., Candeias, S., Pellegrini, T., Abad, A., Tjalve, M., & Dias, M. S. (2014). Automatically recognising European Portuguese children's speech: Pronunciation patterns revealed by an analysis of ASR errors. In *Computational processing of the Portuguese Language* (pp. 1–11).

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., & Guðnason, J. (2017). Building an ASR corpus using Althingi's parliamentary speeches. In *Proc. of INTERSPEECH* (pp. 2163–2167).

Hu, J., Szalay, T., & Ballard, K. J. (2026). Investigating the feasibility of a novel automated speech analysis application for assessing Australian children at risk of speech sound disorders. Abstract and Talk. Speech Pathology Australia Conference. Gold Coast, Queensland, Australia

Jahangir, R., Teh, Y. W., Nweke, H. F., Mujtaba, G., Al-Garadi, M. A., & Ali, I. (2021). Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications, 171*, Article 114591.

Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2021). Using data augmentation and time-scale modification to improve ASR of children's speech in noisy environments. *Applied Sciences, 11*(18), 8420.

Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2022). A formant modification method for improved ASR of children's speech. *Speech Communication, 136*, 98–106.

Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., & Alwan, A. (2005). TBALL data collection: the making of a young children's speech corpus. In *Proc. of INTERSPEECH* (pp. 1581–1584).

Kempe, V., Brooks, P. J., & Gillis, S. (2024). Four decades of open language science: The CHILDES project. *Language Teaching Research Quarterly, 44*, 15–30.

Kulebi, B., Armentano-Oller, C., Rodriguez-Penagos, C., & Villegas, M. (2022). ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proc. of the workshop ParlaCLARIN III within the 13t^{h} Language Resources and Evaluation Conference* (pp. 125–130)

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Liberman, M. Y. (2019). Corpus phonetics. *Annual Review of Linguistics, 5*(1), 91–107.

Liu, H., MacWhinney, B., Fromm, D., & Lanzi, A. (2023). Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research, 66*(7), 2421–2433.

Millasseau, J., Bruggeman, L., Yuen, I., & Demuth, K. (2021). Temporal cues to onset voicing contrasts in Australian English-speaking children. *The Journal of the Acoustical Society of America, 149*(1), 348–356.

OAIC, A.G. (2026). *What is personal information?* Australian Government, Office of the Australian Information Commissioner. https://www.oaic.gov.au/privacy/your-privacy-rights/your-personal-information/what-is-personal-information

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

Paulo, S., & Oliveira, L.C. (2004). Automatic phonetic alignment and its confidence measures. In *Proc. of Advances in natural language processing* (pp. 36–44).

Pérez-Espinosa, H., Martínez-Miranda, J., Espinosa-Curiel, I., Rodríguez-Jacobo, J., Villaseñor-Pineda, L., & Avila-George, H. (2020). IESC-child: an interactive emotional children's speech corpus. *Computer Speech & Language, 59*, 55–74.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication, 45*(1), 89–95.

Potamianos, A., & Narayanan, S. (2003). Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing, 11*(6), 603–616.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N. & Vesely, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 workshop on automatic speech recognition and understanding.*

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., & Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proc. of INTERSPEECH* (pp. 2751–2755).

R Core Team (2025). Vienna, Austria. https://www.R-project.org/

Radha, K., Bansal, M., & Pachori, R. B. (2024). Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review. *Engineering Applications of Artificial Intelligence, 131*, Article 107661.

Rousseau, A., Deléglise, P., & Esteve, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. *LREC,* 3935–3939.

Rumberg, L., Gebauer, C., Ehlert, H., Wallbaum, M., Bornholt, L., Ostermann, J., & Lüdtke, U. (2022). kidsTALC: A corpus of 3-to 11-year-old German children's connected natural speech. In *Proc. of INTERSPEECH* (pp. 5160–5164).

Schultz, B. G., Tarigoppula, V. S. A., Noffs, G., Rojas, S., van der Walt, A., Grayden, D. B., & Vogel, A. P. (2021). Automatic speech recognition in neurodegenerative disease. *International Journal of Speech Technology, 24*(3), 771–779.

Shahin, M., Epps, J., & Ahmed, B. (2025). Phonological level wav2vec2-based mispronunciation detection and diagnosis method. *Speech Communication, 173*, Article 103249.

Shahin, M., Lu, R., Epps, J., & Ahmed, B. (2020). UNSW system description for the shared task on automatic speech recognition for non-native children's speech. In *Proc. of INTERSPEECH* (pp. 265–268).

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

Shobaki, K., Hosom, J.-P., & Cole, R. (2000). The OGI kids' speech corpus and recognizers. *Proc. of ICSLP* (pp. 564–567).

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods, 11*(1), 54–71.

Sobti, R., Guleria, K., & Kadyan, V. (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications, 83*, 1–63.

Sprugnoli, R., Moretti, G., Bentivogli, L., & Giuliani, D. (2017). Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation, 51*, 283–317.

Stevens, M., & Harrington, J. (2016). The phonetic origins of /s/-retraction: Acoustic and perceptual evidence from Australian English. *Journal of Phonetics, 58*, 118–134.

Szalay, T., Benders, T., Cox, F., Proctor, M. (2023). Pre-/l/ vowel change in Australian English *pool-pull*. In *Proc. of the International Congress of Phonetic Sciences* (pp. 2976–2980).

Szalay, T., Nan, Z., Huang, R., Shahin, M., Sirojan, T., Ballard, K. J., & Ahmed, B. (2026). AusKidTalk: Developing transcription guidelines for continuous Australian English child speech. In *Proc. of Language Resources and Evaluation Conference.* (pp. 5794–5804).

Szalay, T., Ratko, L., Shahin, M., Sirojan, T., Ballard, K.J., Cox, F., Ahmed, B. (2022). A semi-automatic workflow for orthographic transcription of a novel speech corpus: A case study of AusKidTalk. R. Billington (Ed.), *Proc. of Australasian International Conference on Speech Science and Technology* (pp. 126–130).

Szalay, T., Shahin, M., Ahmed, B., & Ballard, K. J. (2022). Knowledge of accent differences can be used to predict speech recognition. In *Proc. of INTERSPEECH* (pp.1372–1376).

Szalay, T., Shahin, M., Ballard, K., Ahmed, B. (2022). Training forced aligners on (mis)matched data: the effect of dialect and age. R. Billington (Ed.), *Proc. of the Australasian International Conference on Speech Science and Technology* (pp. 36–40) .

Szalay, T., Shahin, M., Sirojan, T., Nan, Z., Huang, R., Ballard, K.J., Ahmed, B. (2025). AusKidTalk: The benefits of out-of-domain automatic speech processing tools in corpus building. In *Proc. of INTERSPEECH* (pp. 4268-4272).

Virkkunen, A., Rouhe, A., Phan, N., & Kurimo, M. (2023). Finnish parliament ASR corpus: Analysis, benchmarks and statistics. *Language Resources and Evaluation, 57*(4), 1645–1670.

Wagner, M., Tran, D., Togneri, R., Rose, P., Powers, D., Onslow, M. & Ambikairajah, E. (2011). The big Australian speech corpus (the big ASC). *Proc. of the Australasian International Conference on Speech Science and Technology* (pp. 166–170).

Ward, W., Cole, R., & Pradhan, S. (2019). *My science tutor and the MyST corpus*. Carbondale: Boulder Learning Inc.

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication, 140*, 50–70.

Werner-Seidler, A., Maston, K., Calear, A. L., Batterham, P. J., Larsen, M. E., Torok, M., et al. (2023). The future proofing study: Design, methods and baseline characteristics of a prospective cohort study of the mental health of Australian adolescents. *International Journal of Methods in Psychiatric Research, 32*(3), Article e1954.

Yadav, I. C., Kumar, A., Shahnawazuddin, S., & Pradhan, G. (2018). Non-uniform spectral smoothing for robust children's speech recognition. In *Proc. of INTERSPEECH* (pp. 1601–1605).

Authors and Affiliations

Tünde Szalay[1,2,3] · Mostafa Shahin[2] · Tharamkulasingam Sirojan[2] · Zheng Nan[2] · Renata Huang[1,3] · Joanne Arciuli[4] · Elise Baker[5,6,7] · Felicity Cox[3] · Kirrie Ballard[1] · Beena Ahmed[2]

✉ Tünde Szalay
tuende.szalay@sydney.edu.au

Mostafa Shahin
m.shahin@unsw.edu.au

Tharamkulasingam Sirojan
t.sirojan@unsw.edu.au

Zheng Nan
zheng.nan@unsw.edu.au

Renata Huang
renata.huang@mq.edu.au

Joanne Arciuli
joanne.arciuli@flinders.edu.au

Elise Baker
e.baker2@westernsydney.edu.au

Felicity Cox
felicity.cox@mq.edu.au

Kirrie Ballard
kirrie.ballard@sydney.edu.au

Beena Ahmed
beena.ahmed@unsw.edu.au

[1] Discipline of Speech Pathology, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2050, Australia

[2] Department of Electrical Engineering and Telecommunications, Faculty of Engineering, The University of New South Wales, Sydney, NSW 2052, Australia

[3] Department of Linguistics, Faculty of Medicine, Health, and Human Sciences, Macquarie University, Sydney, NSW 2109, Australia

[4] College of Nursing and Health Sciences, Flinders University, Adelaide, SA 5042, Australia

[5] School of Health Sciences, Western Sydney University, Sydney, NSW 2560, Australia

[6] South Western Sydney Local Health District, Western Sydney University, Sydney, NSW 2170, Australia

[7] Transforming early Education And Child Health Research Centre (TeEACH), Western Sydney University, Sydney, NSW 2145, Australia