Author/s:
Ahmed, B;Ballard, KJ;Burnham, D;Sirojan, T;Mehmood, H;Estival, D;Baker, E;Cox, F;Arciuli, J;Benders, T;Demuth, K;Kelly, B;Diskin-Holdaway, C;Shahin, M;Sethu, V;Epps, J;Lee, CB;Ambikairajah, E

Title:
AusKidTalk: An Auditory-Visual Corpus of 3- to 12-Year-Old Australian Children's Speech

Date:
2021

Citation:
Ahmed, B., Ballard, K. J., Burnham, D., Sirojan, T., Mehmood, H., Estival, D., Baker, E., Cox, F., Arciuli, J., Benders, T., Demuth, K., Kelly, B., Diskin-Holdaway, C., Shahin, M., Sethu, V., Epps, J., Lee, C. B. & Ambikairajah, E. (2021). AusKidTalk: An Auditory-Visual Corpus of 3- to 12-Year-Old Australian Children's Speech. Proceedings Interspeech 2021, 6, pp.3680-3684. ISCA. https://doi.org/10.21437/interspeech.2021-2000.

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

AusKidTalk: An Auditory-Visual Corpus of 3- to 12-year-old Australian Children's Speech

Beena Ahmed[1], *Kirrie Ballard*[2], *Denis Burnham*[3], *Tharmakulasingam Sirojan*[1], *Hadi Mehmood*[1], *Dominique Estival*[3], *Elise Baker*[3], *Felicity Cox*[4], *Joanne Arciuli*[5], *Titia Benders*[4], *Katherine Demuth*[4], *Barbara Kelly*[6], *Chloé Diskin-Holdaway*[6], *Mostafa Shahin*[1], *Vidhyasaharan Sethu*[1], *Julien Epps*[1], *Chwee Beng Lee*[3], *Eliathamby Ambikairajah*[1]

[1]University of New South Wales (UNSW), Australia
[2]University of Sydney, Australia
[3]Western Sydney University, Australia
[4]Macquarie University, Australia
[5]Flinders University, Australia
[6]University of Melbourne, Australia

{beena.ahmed, s.tharmakulasingam, h.mehmood, m.shahin, v.sethu}@unsw.edu.au,
{j.epps, e.ambikairajah}@unsw.edu.au, kirrie.ballard@sydney.edu.au,
{denis.burnham, d.estival, e.baker2, chwee.lee}@westernsydney.edu.au,
{felicity.cox, titia.benders, katherine.demuth}@mq.edu.au,
joanne.arciuli@flinders.edu.au,
{barbara.kelly, chloe.diskinholdaway}@unimelb.edu.au

Abstract

Here we present AusKidTalk [1], an audio-visual (AV) corpus of Australian children's speech collected to facilitate the development of speech based technological solutions for children. It builds upon the technology and expertise developed through the collection of an earlier corpus of Australian adult speech, AusTalk [2,3]. This multi-site initiative was established to remedy the dire shortage of children's speech corpora in Australia and around the world that are sufficiently sized to train accurate automated speech processing tools for children. We are collecting ~600 hours of speech from children aged 3-12 years that includes single word and sentence productions as well as narrative and emotional speech. In this paper, we discuss the key requirements for AusKidTalk and how we designed the recording setup and protocol to meet them. We also discuss key findings from our feasibility study of the recording protocol, recording tools, and user interface.

Index Terms: speech corpus, children's speech, Australian English

1. Introduction

Recent advances in speech science have vastly accelerated the use of automatic speech processing applications, such as computational speech analyses and automatic speech recognition (ASR), in mainstream mobile devices and consumer electronics. However, work on ASR for children is sadly lacking, limiting the applications of this technology that are available for children. This is unfortunate as children are potential beneficiaries of various compelling speech and language technology applications, such as remote speech therapy tools, interactive reading tutors, pronunciation coaching, emotion recognition and educational games. The further development of such applications will be made possible by AusKidTalk through the enhanced automated speech processing tools that will result.

Compared to adults, the accuracy of ASR with children's speech is so poor that it is practically unusable [4,5]. One reason for this is that most speech processing systems are developed for and trained on adult speech. Due to the considerable differences between adults' and children's vocal characteristics and word choices, the performance of ASR systems for children's speech remains low [6]. The paucity of publicly available child speech corpora and the small size of such corpora have hindered research [7]. In addition, the strong age-dependent, anatomical, articulatory and language variations across children effectively dilute the available useable data, further contributing to poor results [8]. The difference in English accents across the world also hinders pooling the multiple smaller size children's speech corpora from different countries or backgrounds. Also, children's speech production capability develops over time, with young children's speech likely to contain a range of acceptable errors (e.g., 'three' could be pronounced as 'free', 'fwee', and 'thwee''). In addition, up to 10% of children have a speech sound disorder (i.e., unacceptable errors for their age) [9]. Children with autism can also present with atypical speech errors. For these reasons, a truly representative child speech corpus needs to contain both typical and atypical speech samples across a wide age range.

Currently, there are less than 20 child speech corpora worldwide, of which only three are sufficiently large for development of speech processing systems such as ASR. Additionally, all, including these three, have used problem-specific protocols with limited tasks, and none is fully annotated [7]. Proposed solutions include harvesting a large amount of child speech from YouTube Kids videos [10] and transforming adult speech [11]. Neither are ideal and, in addition, there are no sufficiently large, public or proprietary Australian child speech corpora. These shortcomings will be remedied in AusKidTalk.

http://dx.doi.org/10.21437/Interspeech.2021-2000

The lack of research on automated speech processing tools for children may be attributed to the fact that child speech, especially that of younger children, is relatively difficult to collect and analyze. Based on our experience with the adult AusTalk, and through the involvement of an interdisciplinary team incorporating engineers, programmers, linguists, speech pathologists, child psychologists and psycholinguists, AusKidTalk has developed solutions to the many problems involved in collecting child data. The aims of AusKidTalk are

1. To collect AV speech from Australian children using state-of-the art infrastructure,
2. To transcribe all the collected acoustic speech data orthographically and annotate the scripted speech at the phoneme level.
3. To enable access via a public research data repository.

2. Data Collection

To meet these specified aims, we set out the requirements below for the AusKidTalk recording equipment setup and protocol:

1. The protocol should elicit different utterance types, e.g. single words, sentences, narratives, emotional speech, to facilitate research into and the development of a broad range of speech processing applications.
2. Samples from a wide range of ages and utterance types should be collected to develop a solid theoretical understanding of the structure and mechanisms of Australian children's speech needed to support development of children's speech-based applications.
3. Samples of both typically developing and disordered speech should be collected to provide a truly representative child speech corpus and facilitate work on automated tools to benefit children with disordered speech.
4. The recording setup should be child-friendly and be able to maintain engagement with children for 1-2 hours.
5. The recording setup and protocol should facilitate automated pre-processing and segmentation to reduce the manual phoneme-level annotation time.
6. The resulting corpus should be sufficiently large to train speech processing applications for children such as ASR, pronunciation assessment, speech disorder detection and emotion recognition algorithms.

To achieve these requirements in AusKidTalk, we developed infrastructure to collect audio visual (AV) recordings of Australian English (AusE) typically- developing and disordered speech from 750 children from 3 to 12 years, equally distributed by age. Each child participates in a structured 90-120 min recording session, with breaks as necessary depending on the child's age and attention. Based on our combined experience and expertise, we have developed a child-friendly recording setup in which AV recordings are collected from the children minimally invasively. Game-based activities are used to elicit and record speech samples thus maintaining better engagement during the session.

In the recording sessions, children participate in several activities designed to elicit different types of speech (e.g. single words, sentences, narratives) needed to train speech processing applications as well as to better understand children's speech. Collectively these activities elicit 30 mins of continuous clean usable speech from each 3-5 year old; 45 mins from 6-8 year olds; and 60 mins from 9-12 year olds. The total of approximately 600 hours will provide a corpus size needed to meet the development needs for speech processing applications such as children's ASR systems with low WERs (word error rates), automated speech to text annotation/transcription and mispronunciation detection systems of sufficient accuracy, as well as speech disorder classification and atypical speech identification systems acceptable clinically.

2.1. Participants

Institutional Review Board (IRB) approval was obtained from the Ethics Committee of all involved universities (lead UNSW) and participants are being recruited by advertisements in social media, online parent group websites and directories, primary schools, school/community newsletters, university media releases and flyers sent to participant databases maintained by the researchers at each university. Prior to enrolment, carers provide electronic consent on a secured website and complete an online demographic survey to verify that children meet the eligibility criteria: that they are native AusE speakers, defined here as a child born and only educated in Australia with at least one parent who completed all their primary and secondary schooling in Australia. All the recordings are anonymized to protect the identity of the children.

Of the 750 children, recordings are being collected from 700 typically developing (TD) children and 50 children with speech sound disorder (SD). Half of the SD group have no other developmental diagnosis and half have a diagnosis of autism spectrum disorder (ASD). The SD children have, according to parent-report, a speech disorder in the absence of any problems with hearing or vision, plus either (a) no other history of developmental disorder or (b) a prior diagnosis of autism spectrum disorder. Both TD and SD groups are between the ages of 3 – 12 years, except for the ASD group. Children in the ASD group are aged 6 to 12 years as the diagnosis is often not as reliable/confirmed until that age.

2.2. Recording Protocol

Speech is being collected in a variety of activities designed to reflect the range of uses in child communication and different skill levels (e.g. constrained speech, such as numerals, through to unconstrained speech). As set out below, there are five tasks in the recording protocol, which combined will result in a corpus with the full set of phonemes, consonant clusters, vowel-consonant pairs, lexical stress patterns and emotional range required to research a range of automated speech analysis tools and better understand the speech and language abilities of children. Video recordings of the whole session are also made to support manual annotation of the children's speech (e.g. manual phoneme-level annotation of speech in this population who make developmental errors) and to produce reference data for future research on gesture.

2.2.1. Single Words

This activity (the Speech Test for Australian Children – STAC [12]) is designed to gather recordings of each child producing target words, ideally spontaneously, otherwise via imitation of a standard pre-recorded prompt. The task comprises each child producing 117 single words (115 test words and 2 practice words), numbers from 1 to 10, and 12 two-word 'counting' phrases (e.g. one egg, two eggs, three eggs, four eggs). The STAC contains words:

- varying in syllable length (50% monosyllables, 20% disyllables, 30% polysyllables),

- varying in lexical stress patterns (trochaic and iambic),
- that collectively sample all AusE consonants at least once in phonotactically permissible word-initial, -medial and -final positions,
- that collectively sample all AusE vowels, and
- with a range of word-initial and -final consonant clusters.

The children's responses will not only provide data for accurate modelling of phoneme and word productions but will also ensure a standard set of productions for use in a range of analyses to determine characteristics of AusE across children of differing ages and socio-demographic backgrounds. The children's responses will also provide valuable information for studying AusE-speaking children's speech acquisition for consonants, vowels, word length, and lexical stress.

2.2.2. Sentence Repetition

This activity is designed to explore issues of speech planning as a function of prosodic complexity within utterances and the effects of prosody and utterance length on the production/omission of function words. The child is instructed to listen to 36 pre-recorded sentences of varying word-lengths, intonational patterns (e.g. question vs statement), and grammatical structures, and then repeat what they hear. These responses will provide a standard set of children's productions of connected speech from a range of linguistic contexts that complement the collected single-word productions.

2.2.3. Story Telling

In this activity, the child tells a story using images presented on a tablet. The use of pictures reduces verbal scaffolding that might be provided by the researcher and encourages the child to generate a story based on the standard set of picture stimuli. Researchers restrict their verbal feedback to comments that will not influence the child's form of expression (i.e. by avoiding prompts that might lead to a particular choice of verb tense, aspectual marking, or perspective). The activity uses a picture sequence from Doggy Cartoons' story of a boy and a dinosaur [13]. Video collected in this activity captures gesture as well as facial and verbal expressions. The child's responses will provide connected speech for use in automated speech analysis, in addition to insight into the development of children's narrative abilities and their use of increasingly complex syntactic and discourse structures in telling narratives.

2.2.4. Emotion Elicitation

The objective of this activity is to elicit speech with emotional content, noting the paucity of available data with emotional children's speech (FAU Aibo [14] is perhaps the only major example to date), particularly from different ages. To achieve this, the child first watches two short movie clips (one happy, one sad) and then is asked a series of questions about each (e.g., to describe each scene, how it makes them feel and why). The child is encouraged to engage with the scenes and produce responses to each question that are at least a sentence long. The collected responses will assist in characterizing emotional speech in children of different ages and the development of emotion recognition tools for children.

2.2.5. Nonsense Words

This task is concerned with examining the mechanisms of children's speech production, specifically phonological structure, without the influence of semantics, and phonological working memory. The child is asked to repeat (imitate) 40 nonsense words from the Children's Nonword Repetition test [15] (e.g. ballop, pristoractional). Children's productions on this task will provide valuable developmental data on their ability to repeat words they have not heard before--an important skill that children need to develop to grow their vocabulary. The productions will also assist in the development of automated speech-based systems that require children to repeat words they may or may not know. Children's production of nonwords will also provide insight into their underlying speech processing skills [15]. Accuracy of imitation is considered a clinically valuable marker for a range of speech and language disorders [16] and thus beneficial for the development of speech disorder assessment systems. The collected responses will assist in describing the speech and language abilities of the children's samples to the AusKid Talk corpus.

2.3. Data Collection Equipment

To collect the recordings, we upgraded existing equipment used for AusTalk [2,3], the precursor of this project in which a corpus of Australian adult speech was collected. The upgrades included new PCs and external hard drives with more current models for each recording station, child-friendly headsets, chairs, tablets to present the children with the protocol prompts via game-like applications and a data server (see Table 1), and child-friendly chairs.

Table 1: *Data Collection Equipment*

Equipment	
Audio mixer	Child headset
Recording computer	Far field mic on table
External hard drive	2x stereo mics
HD Stream Webcam	2x 10.4 in Android tablets
17 in monitor	Central data server
Bluetooth speaker	

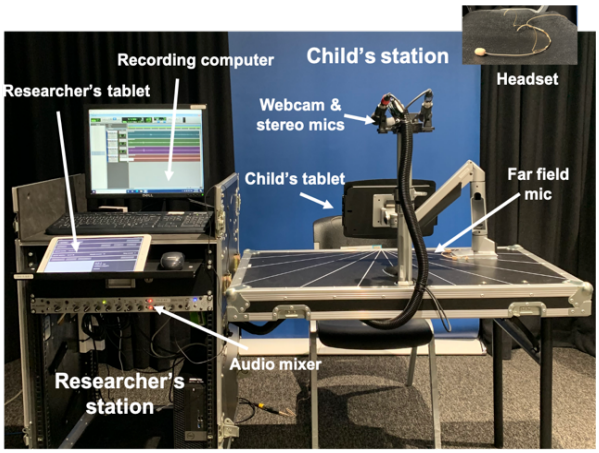

Figure 1: *Setup of data collection equipment as in Table 1*

A child-friendly, flexible and light, microphone-only headset is used to keep the recording setup minimally invasive. The child's tablet is mounted on an adjustable arm mounted on the table to accommodate different children's heights as shown in Figure 1. This helps to limit the interaction the child makes with the tablet allowing them instead to focus on the recording task and only press buttons on the tablet when prompted.

As shown in Figure 1, all microphones are fed into the audio mixer, which is connected to a main recording computer. The researcher controls the recording of the audio and video from this computer. This includes calibrating the microphone, verifying audio quality, configuring the video camera and starting/stopping the recording. All recordings are saved locally in the local drive and on an external hard drive and then backed up automatically on the project-specific server established at UNSW, as well as in the UNSW Data Archive. The data will be transferred to an online repository for long-term storage and research community access on completion.

2.4. User Interfaces

Our previous experience in recording speech from children [17] has shown that actively engaging children during the sessions facilitates optimal data collection. Accordingly, we developed custom game-based tablet applications to deliver visual and pre-recorded audio stimuli and prompts through an engaging set of tasks interspersed with non-verbal incentive games to maintain the child's enthusiasm. The researcher has manual control via audio-record buttons for repetition attempts where necessary. Except for the task where the child is required to imitate a nonsense word, options are given for spontaneous or imitated responses (using controlled pre-recorded stimuli) depending on individual age/ability. We also created additional interactive, non-speaking tablet games to engage the children between the different activities and maintain their enthusiasm.

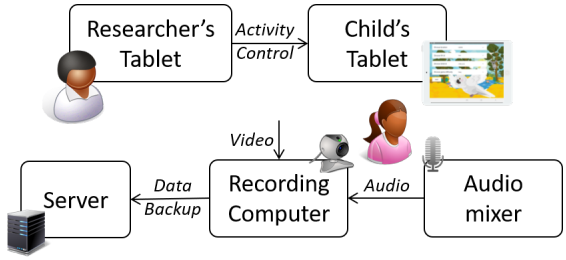

Figure 2: *Data Collection setup. The researcher controls the activity of the child's tablet via their tablet. Audio from all the microphones is fed to the mixer and sent to a local computer with the webcam video for back up on a server.*

The presentation of the activities on the child's tablet is controlled by the researcher through a separate tablet linked over the internet (see Figure 2). Via their tablet application, the researcher controls which activity is presented to the child, whether to repeat, skip or go back to a stimulus in the activity. For the single word and sentence activities, the researcher also marks whether the child's response is correct (i.e., an attempt of the intended response regardless of speech accuracy) or incorrect (defined as no attempt or production of a different response to the intended response). If a child is losing interest in an activity, the researcher can pause and present the child with one of the non-speaking games to re-engage them; or if a child is unable to continue, they can also exit that activity.

To facilitate automated speech segmentation and annotation of the recordings, timestamps are generated each time the researcher interacts with their tablet application. This includes timestamps for each time an activity starts, each time a stimulus is presented to the child (both offset and onset times), each time the child responds to a stimulus, each time the activity ends and each time the child re-attempts a certain stimulus. These timestamps are then used to segment the speech into individual files with names based on the stimuli.

3. Recording Status

Feasibility of the recording protocol, recording tools, and user interface was conducted with 231 children (128M, 103F). 84% had typically developing speech, 13% speech sound disorder, and 4% autism. The majority of the children completed all five tasks in the recording protocol. Some tasks were more challenging for some 3-year-olds. For instance, although all 3-year-olds attempted Task 1 (production of single words), 20% of 3-year-olds skipped Task 2 (repeating sentences), 4% skipped Task 3 (telling a story), 16% skipped Task 4 (responding to questions about one happy and one sad short movie clip), and 8% skipped Task 5 (imitation of increasingly long nonsense words). This was not unexpected, given that 3-year-olds are still learning to communicate and developing their ability to produce sentences, tell stories, and use their phonological working memory to learn new words.

Regarding the user interface, one important consideration was to select appropriate and engaging visual stimuli (pictures and videos) for children from 3- to 12-years. In addition, for the emotion elicitation task, the lack of literature guidance made the choice of videos that appropriately balanced scariness (for younger children) with childness (for older children) nontrivial. The inclusion of the non-speaking tablet games during and between tasks helped to facilitate engagement and completion of the recording protocol. The use of two separate tablets in the recording setup ensured children could engage with the games while the researcher managed task completion and set up.

Table 2 below presents the current recording status.

Table 2: *Recordings collected to date*

	Male	Female	% completed
3-5 yrs	68	62	58
6-8 yrs	66	51	52
9-12 yrs	66	36	34
Disordered	46	11	100

4. Conclusions

AusKidTalk will be the first Australian children's speech corpus to meet the demands of modern speech science. It will provide speech samples from 750 children (i) across a diverse age range (3 to12 years) to capture both between-speaker and across-age variation in Australian speech and language, (ii) across different purpose-specific speech activities, some tightly controlled to capture within-speaker and between-context variability. AusKidTalk will enable the development of automated speech algorithms for child-specific applications, benefiting Australia's regional and remote communities.

Here, we have shown that it is feasible to collect speech consistently from across the full age range of 3- to 12- year old children. Key factors responsible for this were the interactive, child-friendly recording setup developed, the researcher control of the activities and the variety of activities used to keep the children engaged. The recording setup has been designed to facilitate automated annotation of the scripted data enabling the provision of phonemic annotation, which few corpora provide.

5. Acknowledgements

We would like to thank Margaret Ryan for her help in data processing. This project was supported by the Australian Research Council (LE190100187) as well as the University of New South Wales, The University of Sydney, Western Sydney University, Macquarie University and The University of Melbourne.

6. References

[1] http://www.auskidtalk.edu.au/

[2] https://austalk.edu.au/

[3] Burnham, D. et al (2011). Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box, In Proc. of Interspeech, 2011, 841-844.

[4] Russell, M., & D'Arcy, S. (2007). Challenges for computer recognition of children's speech. In Workshop on Speech and Language Technology in Education.

[5] Elenius, D., & Blomberg, M. (2005). Adaptation and normalization experiments in speech recognition for 4 to 8 year old children. In Proc. of Interspeech.

[6] Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. In Proc. of Interspeech

[7] Chen, N.F., Tong, R., Wee, D., Lee, P.X., Ma, B. and Li, H., 2016. SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese. In Proc. of Interspeech (pp. 1545-1549).

[8] Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America, 105(3), 1455-1468.

[9] McGregor, K. K. (2020). How We Fail Children With Developmental Language Disorder. Language, Speech, and Hearing Services in Schools, 51(4), 981-992. https://doi.org/doi:10.1044/2020_LSHSS-20-00003

[10] Baker, E., Cox, F., Arciuli, J., & Ballard, K. (2019). Speech Test for Australian Children (STAC). Sydney, Australia

[11] Liao, H., Pundak, G., Siohan, O., Carroll, M., Coccaro, N., Jiang, Q.M., Sainath, T.N., Senior, A., Beaufays, F. and Bacchiani, M., (2015). Large vocabulary automatic speech recognition for children. In Proc. of Interspeech (pp.1611-1615).

[12] Serizel R, Giuliani D. (2014). Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In Proc of IEEE Spoken Language Technology Workshop (pp. 135-140).

[13] Doggy Dog, https://www.youtube.com/watch?v=DTfv8y05fj8

[14] Steidl, S. (2009). Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, Logos Verlag, Berlin, 2009.

[15] Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. Memory, 2(2), 103-127. https://doi.org/10.1080/09658219408258940

[16] Pigdon, L., Willmott, C., Reilly, S., Conti-Ramsden, G., & Morgan, A. T. (2020). What predicts nonword repetition performance? Child Neuropsychology, 26(4), 518-533. https://doi.org/10.1080/09297049.2019.1674799

[17] Ahmed, B., Monroe, P., Hair, A., Tan, C.T., Gutierrez-Osuna, R. and Ballard, K.J., (2018). Speech-driven mobile games for speech therapy: User experiences and feasibility. International journal of speech-language pathology, 20(6), pp.644-658.