Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box

Denis Burnham[1], *Dominique Estival*[1], *Steven Fazio*[1], *Jette Viethen*[2], *Felicity Cox*[2], *Robert Dale*[2], *Steve Cassidy*[2], *Julien Epps*[3], *Roberto Togneri*[4], *Michael Wagner*[5], *Yuko Kinoshita*[5], *Roland Göcke*[5], *Joanne Arciuli*[6], *Marc Onslow*[6], *Trent Lewis*[7], *Andy Butcher*[7], *John Hajek*[8]

[1]MARCS Auditory Laboratories, University of Western Sydney, Australia
[2]Macquarie University, Australia
[3]University of New South Wales, Australia
[4]University of Western Australia, Australia
[5]University of Canberra, Australia
[6]University of Sydney, Australia
[7] Flinders University, Australia
[8]University of Melbourne, Australia

{d.burnham,d.estival,s.fazio}@uws.edu.au, johnth@unimelb.edu.au, jette@gmail.com,
{felicity.cox,robert.dale,steve.cassidy}@mq.edu.au, j.epps@unsw.edu.au
roberto.togneri@uwa.edu.au, {michael.wagner,yuko.kinoshita,roland.goecke}@canberra.edu.au
{joanne.arciuli,mark.onslow}@sydney.edu.au, {trent.lewis,andy.butcher}@flinders.edu.au

Abstract

The Big Australian Speech Corpus project incorporates the strategic goals of 30 Chief Investigators from various speech science areas. Speech from 1000 geographically and socially diverse speakers is being recorded using a uniform and automated protocol plus standardized hardware and software to produce a widely applicable and extensible database – AusTalk. Here we describe the project's major components and organization; share the lessons learnt from difficulties and challenges; and present the results achieved so far.
Index Terms: speech corpus, AV data, Australian English.

1. Introduction

1.1. Background

In 2009, a group of 13 institutions (see Acknowledgements) and 30 Chief Investigators received funding from the Australian Research Council for the Big ASC (Big Australian Speech Corpus) [1], now known to the public as AusTalk [2].

Australian English (AE) is a regional dialect, which, like all spoken dialects, contains some variation. The Macquarie Dictionary of Australian English (first ed. 1981) recognised 3 main varieties of AE – broad, general and cultivated, considered to be loosely related to social group membership, but regional variations were not considered sufficiently salient to warrant discussion. Today we can identify a broader range of AE varieties including ethno-cultural dialects which in more recent years have become increasingly evident and proudly spoken [3]. While certain varieties have been investigated to some extent [4] and subtle regional variations are now recognised [5], much remains unknown. The only publicly available Australian speech corpus is the 14-year-old Australian National Database of Spoken Language (ANDOSL) [6], audio-only low-fidelity single-session corpus of 108 speakers on a limited number of tasks. The main purpose of AusTalk is to provide data for projects charting the extent, degree, and details of social, regional and ethno-cultural variations in AE, but it will also many other research projects and applications. Five strengths of AusTalk are as follows: (1) The cornerstone of any corpus is its *variability*; AusTalk is no exception, with 1000 speakers from 17 different locations. (2) AusTalk will benefit from *standardisation*, due to standard recording hardware and software, corpus protocol and annotations, and (3) the greatest possible degree of *automation.* (4) Moreover, with project partners from the full range of speech science research areas, the Big ASC is highly *collaborative* and *widely applicable.* (5) Finally, AusTalk will establish an *extensible system* of 12 identical recording stations spread across Australia, with a central storage, access and annotation system, from which the corpus will be freely available for research purposes.

1.2. Scope of the Big ASC project

The initial AusTalk corpus will comprise 1000 AE speakers, all having completed primary and secondary school in Australia, a criterion ensuring inclusion of a range of speakers from various cultural backgrounds. With 3 1-hour sessions each, this totals 3000 hours of high quality audio-visual data.

1.2.1. Variability

The 3 1-hour-sessions are recorded at intervals of at least one week to capture potential *variability over time*, while *geographical variability* is guaranteed by recording at 17 locations, covering all capital cities of Australian states and territories and several regional centres. Speakers from a range of social spheres (*social variability*) will be included due to wide advertising and high visibility (< 2 months after the well-publicised launch of the project, more than 700 participants had registered on the website: http://austalk.edu.au). The protocol ensures a variety of speech content from a range of tasks: 4 Read Speech and 4 Spontaneous Speech tasks.

1.2.2. Standardisation

Standardisation across recording locations is made possible by 12 identical portable, self-contained and cost-effective

recording stations (Black Boxes); the fixed protocol that ensures systematic data collection across locations; and a 2-day central training session for all Research Assistants (RAs) to ensure procedural uniformity across locations.

1.2.3. Automation

Standardisation is maximised by the high degree of automation, mainly provided by the Standard Speech Collection Protocol (SSCP) software which minimises risk of error by either RAs or speakers during recording. The SSCP also automates data acquisition by starting, stopping and synchronising hardware; data quality monitoring and checking; file labelling and data uploading .

2. Components of the Big ASC

2.1. Hardware and Equipment: the Black Box

The Black Box (see Figure 1) is *self-contained* and composed mostly of off-the-shelf elements (except for the tailor-made sync cable & camera mounts) making it *economical* (the stereo cameras are the most expensive items), *transportable*, and easily *replicable*. A complete list of equipment follows.

Figure 1: Black Box

1 Mixer Rack Workstation: the 'Black Box' for storing and transporting items; unpacks into 2 tables & computer rack
2 Capture Computer: PC for protocol display and recording.
3 Audio recording device: M-Audio FastTrack Ultra8R.
4 Head worn mic (x2): AudioTechnica AT892c.
5 AT8539 Phantom Power/XLR adapter to connect mic.
6 Far-Field mic: Shure MX391/O. On table, ~ 60cm from speaker.
7 Stereo mics (x2): Behringer C-2. On table, ~60 cm from speaker, to record hands-free voice interaction conditions.
8 Stereo Cameras BumbleBee2 (x2). Mounted ~50cm from speaker. Dual bus firewire card.
9 Custom -made GPIO to audio Sync Cable: for audio/video synchronisation: camera sends a strobe signal out to the M-Audio DAQ to record a waveform.
10 17inch Monitors 4:3 (x2): Dell E170S 17 inch Flat Panel Monitor. To display prompts to speaker and for RA.
11 Monitor arm / stand: Atdec Visidec Focus MICRO LCD Single Arm, VF-M. To hold monitor and camera.
12 Tripod mount for camera (x3): Manfrotto 700RC2 tripod
13 External hard drive: Samsung STORY Station 2TB.
14 Light Meter: Lux & Fc Light Meter (DSE).
15 Operator Head Phones: KOSS UR-20, for the RA.
16 Partition for Map Task: 400x500 3mm black acrylic board
17 Pull-up backdrop (x2) to provide uniform background.
18 Chairs (x2) to ensure standardisation of video capture.

2.2. AusTalk Collection Protocol

The SSCP guarantees collection of similar data across all the speakers. The Calibration, Yes/No and the first 3 Read Speech tasks are repeated across all 3 sessions, with one Spontaneous Speech task per session (Session: Interview; Session 2: Re-told Story; Session 3: Map Task).

2.2.1. Calibration

Room noise ('no speech') is recorded at the start of each session for 1 min to capture all aspects of noise in the room, e.g., air-conditioning, workstations, plus any external noise. Stereo microphones ensure both additive and reverberant noise behaviours reflecting the room acoustics.

2.2.2. Read Speech

Digits: A simple isolated digit task (12 4-digit sequences) is designed for speaker verification applications. Sequences were selected to ensure each of the 10 digits occurs at least once in each serial position, to capture any co-articulatory variations and to provide sufficient acoustic-phonetic variation.

Individual Words: In this task speakers read words that are randomly presented via computer screen. The product is citation form productions of 322 words of 3 word types:

A. 77 monosyllabic words comprising the stressed vowels of AE in the standard hVd, hVt, hV, hVl, hVn contexts
B. 167 words to address specific AE phonetic features
C. 68 polysyllabic words to sample variations in lexical stress

Set A will allow carefully controlled acoustic examination of vowel systems across speakers and dialects and is required for socio-phonetic and forensic work on variation in AE. The standard format (used in phonetic vowel studies globally) provides a data set that can be easily and validly compared with other similar data sets. Set B [5] provides added scope for comparing across contexts not represented in the standard format and allows capture of contextually variable realisation of consonants. Set C [7] will allow fine-grained acoustic analyses of stress contrastivity within and across words.

Sentences: A set of 58 constructed sentences is designed to elicit connected speech in a standard format. 50 sentences were derived from the phonetically-rich list of the more limited 1995 ANDOSL [6], based on the SCRIBE sentences) to sample all vowels and consonants in a range of connected permutations with varying prosodic characteristics; this will allow direct comparison of various connected speech processes (CSPs) across the Austalk corpus and with previously collected AE speech, including migrant varieties. The remaining 8 sentences were designed to elicit additional processes, e.g., AE diphthongal features which have been used to define AE speaker groups.

Read Story: This task has a dual purpose: first to provide a 'launch pad' for a spontaneous narrative, secondly to provide the possibility for comparison between speech styles within and between speakers. The criteria for selection of the text were that it must (1) include a coherent story; (2) be sufficiently complex to produce a re-telling of reasonable length (> 60 s) but not so complex as to challenge memory; (3) be in reasonably informal language; (4) be culturally appropriate. The story is a version of the tried and tested 'Arthur the Rat'[8], a phonetically balanced text that has been Australianised to include more local lexical and grammatical features and important CSPs for AE, and which also samples prosodic parameters, including pausing and breathing.

2.2.3. Spontaneous Speech

Interview: This captures spontaneous, engaged, narrative talk, i.e. 'story telling' in the vernacular style [9]. While this style of speech is hard to capture in formal interviews, it can be achieved with skilful empathetic interviewers and particular attention to facilitating elicitation conditions. Different topics are suggested for discussion, subject to speakers' preferences.
Re-told Story: Speakers are asked to re-tell the 'Arthur the Rat' story they have just read, to allow within- and between-speaker speech style comparison. Previous work suggests that for a re-told text, the duration is about 15% shorter than the original, but with a wide range – 45% shorter to 30% longer.
Map Task: This is a data gathering game to collect spontaneous speech in a dialogic setting. Each of the 2 participants is given a map of the same environment with a number of landmarks. One map also shows a route that winds between the landmarks from a start to a finish point. The task for the instruction giver (IG) is to describe this route to the other participant (the instruction follower, IF) who must draw the route onto the other map. Landmark discrepancies between the maps (different visual appearance or label, or absence) are designed to encourage IF-IG negotiation. The aims of the AusTalk Map Task are to (1) sample a number of phonological segmental and CSPs in truly spontaneous speech which can be compared to the more formal speaking tasks; and (2) collect a corpus of object descriptions that allows study of various discourse phenomena, e.g., negotiation over reference. Accordingly, the AusTalk maps contain 2 types of landmarks: PHON (labels forcing participants to utter the exact words and CSPs of interest) and REG (depictions of buildings that can be distinguished visually but are not given labels).
Yes/No: This component consists of expressions meaning 'Yes' or 'No' and all their variations, useful for forensic applications. The target expressions are elicited from the necessary conversations occurring naturally during the sessions and are recorded in all 3 sessions to capture within-speaker variability. The Research Assistants use a list of prepared questions at the beginning and end of each session.

2.3. Data Storage and Annotation

Two important goals of the project are to make AusTalk widely available and to allow future contributions, either further data or further annotation. Audio and video data will be stored on a web accessible server, with corpus meta-data and annotations stored in the DADA annotation store [10]. The DADA server can support import/export of annotation data in formats supported by many annotation and analysis tools.

Initial annotation is limited to word segmentation for the Read Speech and to transcription aligned at the phrase or sentence level for Spontaneous Speech. Forced alignment will be used where appropriate to enhance manual annotation. New annotations, e.g. detailed phonetic transcription, can be contributed by project partners or other researchers and integrated into the existing annotation store.

3. Variations to the AusTalk Core

3.1. Australian Aboriginal English

Australian Aboriginal English (AAE) is spoken by the majority of Australia's almost 0.5 million Aboriginal population. Recordings are to be made in Darwin and Alice Springs, with 24 Aboriginal and 24 non-Aboriginal speakers in each location. There is no intention to exclude Aboriginal English speakers at other centres where it is expected that a small number will also be recorded. The recording of creole varieties and traditional indigenous languages is postponed for a separate future project using appropriate elicitation methods.

3.2. Emotions in speech

AusTalk is the first large-scale corpus of Australian speech to include a significant component of emotional speech. This component is collected at 1 site (UNSW), where it replaces the Spontaneous Speech Tasks for 36 speakers. In all 3 sessions, they are asked to view a number of separate randomly-ordered blocks of affective pictures from the International Affective Picture System [11] for 30s to induce different arousal patterns. The protocol is identical between sessions, but the pictures differ. Participants are asked to verbalise their thoughts, feelings and memories about each picture, and in the final session are also asked to (1) view and respond to a small number of video clips from well-known TV shows or films (e.g. Bill Cosby or 'Silence of the Lambs'); and (2) relate a time when they felt, e.g., sad, in order to elicit natural conversational speech coloured by real emotional experiences.

3.3. Stuttering

AusTalk will establish the first extensive normative data (16 speakers recorded with the same standard protocol) to guide the treatment of stuttering in Australia. For example, treatment for chronic stuttering sacrifices speech naturalness to some extent in order to control stuttering, but normative data for the usual measures of speech naturalness are not available. Additionally, many normal disfluencies of speech resemble stuttering, and it is not clear what stuttering severity scores are given to normal speakers.

4. Difficulties and challenges

4.1. Synchronisation

Capturing from up to 3 microphones and 2 stereo cameras at the same time poses potential data synchronisation problems, for which we have both hardware and software solutions: the audio devices are hardware synchronized and the 2 stereo cameras are synchronized via the PointGrey MultiSync software. Inspired by [12], the audio and video streams are synchronized via a 'sync' cable directly connecting the main camera to the M-Audio device as an audio input. The output signal (strobe pulse) can be fired for every frame captured by the camera. Thus, we record 4 synchronous 'audio' inputs with the audio device and, by starting the audio recording first, align the audio and video frames with high accuracy during subsequent analysis.

4.2. Size of video data: compression

Audio data can be stored in high quality lossless format with each channel stored individually, as this constitutes a relatively minor storage requirement. On the other hand, storage requirements for the video data are immense: raw data from the stereo camera is 640x480x16 bits per frame and at 48fps over 3 hours equates to ~300GB for one speaker. Multiplied by 1000 speakers, storage becomes infeasible for raw video data. The most straightforward solution would be to compress all video but this would introduce artifacts which cannot be fixed, so data would be lost forever. Thus we plan to:

1. Store strategic corpus components in full raw format.
2. Store the cropped face region (30-50% of the video frame)
3. Apply feature extraction algorithms (PointGrey depth map, Active Appearance Models, etc) to the raw data.

5. Collaboration and Wider Extension

In the initial stage, speech is only recorded from adult AE speakers. However, with 12 identical Black Boxes available around Australia, and a central storage system, the corpus can be easily augmented, e.g. by adding child speech or particular ethno-cultural varieties of AE. Given the number of institutions (N=13) and the degree to which the 30 investigators' interests and skills have been incorporated, the corpus will be appropriate for applications and future research projects in Cognitive Science and Psycholinguistics, Engineering and Spoken Language Processing, Language Technology and Computer Science, Phonetics and Linguistics, Forensic Speech Science, and Speech Pathology.

5.1. Robust ASR and Speech in Noise (SpIN)

The protocol design included consideration of realistic noise capture in a typical office environment and a small isolated digit recognition task for speaker identification. The 'no speech' noise samples allow noise to be added to the speech recordings in post-processing at different signal-to-noise ratios, to create sufficiently noisy hostile environments for robust speech recognition research.

5.2. Audio-Visual ASR

Progress in AV speech processing technology and the availability of high quality low-cost cameras have made it possible to record an AV speech corpus without great difficulty. We designed a stereo camera recording setup which opens up a number of interesting avenues for multimodal corpus analysis. Not only can we investigate the correlations between audio and video data for AV ASR, but we can do so using 3D face model data, thus enabling improved accuracy for non-frontal faces commonly found in naturally occurring speech. The AusTalk video data can also be used for research in other applications, e.g. face recognition and liveness detection in biometrics, facial expression recognition in affective sensing, and face synthesis in computer graphics and embodied conversational agents.

5.3. Speaker Verification

The isolated digit corpus component can be used to calibrate speech recognition systems and for speaker verification applications, most notably in the still-common use of 4-digits PINs. This component will greatly benefit speaker verification systems developed from limited data for enrolment/training.

5.4. Forensic Applications

For applications such as automatic speech recognition, speaker verification and text-to-speech synthesis, corpora must represent the statistical distribution of speech in the general population as closely as possible. To limit corpus variability in this initial phase, data collection was restricted to the (large) subset of the Australian population that has had all schooling in Australia, a somewhat problematic restriction for forensic applications where foreign-accented speech is encountered reasonably frequently. Hence, one important extension of the corpus will be the collection of foreign-accented English that is spoken by sizeable sections of the Australian population. Nevertheless, considering that both incriminating speech and police interviews typically contain many 'Yes/No' style responses, AusTalk will be valuable in testing forensic speaker recognition approaches and conducting real-world casework. We will also collaborate with a recently-awarded ARC Linkage Project on reliable forensic voice comparison [13].

6. Conclusions

Together, the portable economical equipment and the diversity of the protocol ensures that AusTalk is the first corpus of its kind: speech is captured in a realistic noise environment, and augmented by the addition of facial expression and depth information through stereo cameras. It is the epitome of a modern speech corpus: a variety of AV data, using state-of-the-art recording equipment under standardised automated conditions to yield a large database applicable to a range of research areas and extensible beyond the life of the project – while the corpus covers the urban/regional distinction, and social and geographical variation in AE, but it can later be extended to other age groups and other varieties. By providing infrastructure for future speech science research in Australia, the Big ASC will spawn technologies for real life applications.

7. Acknowledgements

We acknowledge financial and/or in-kind assistance of the Australian Research Council (LE100100211), ASSTA; the Universities of Western Sydney, Canberra, Melbourne, NSW, Queensland, Sydney, Tasmania and Western Australia; Macquarie, Australian National, and Flinders Universities; and the Max Planck Institute for Psycholinguistics, Nijmegen.

8. References

[1] Burnham, D., et al., *A blueprint for a comprehensive Australian English auditory-visual speech corpus*, in *2008 HCSNet Workshop on Designing the Australian National Corpus*. 2009, Somerville, MA, USA: Cascadilla Proceedings Project: Sydney. p. 96-107.

[2] Wagner, M., et al., *The Big Australian Speech Corpus (The Big ASC)*, in *13th Australasian International Conference on Speech Science and Technology*. 2010, ASSTA: Melbourne. p. 166-170.

[3] Cox, F. and S. Palethorpe, *Timing differences in the VC rhyme of standard Australian English and Lebanese Australian English (submitted)*, in *ICPhS XVII*. 2011.

[4] Clyne, M., E. Eisikovits, and L. Tollfree, *Ethnic varieties of Australian English*, in *Varieties of English Around the World: English in Australia*, D. Blair and P. Collins, Editors. 2001, Benjamins: Amsterdam. p. 223-238.

[5] Cox, F. and S. Palethorpe, *Illustration of the I P A: Australian English.* JIPA 2007. **37**: p. 341-350.

[6] Vonwiller, J., et al., *Speaker and Material Selection for the Australian National Database of Spoken Language.* Journal of Quantitative Linguistics, 1995. **3**: p. 177-211.

[7] Carter, A. and C. Clopper, *Prosodic effects on word reduction.* . Language & Speech, 2002. **45**(4): p. 321-353.

[8] MacMahon, M., *The woman behind Arthur.* JIPA. 1991. **21**: p. 29-31.

[9] Labov, W., *Sociolinguistic patterns*. 1978, Oxford: Blackwell.

[10] Cassidy, S. and T. Johnston, *Ingesting the Auslan Corpus into the DADA Annotation Store*, in *Third Linguistic Annotation Workshop (LAW III)*. 2009: Singapore.

[11] Lang, P.J., M.M. Bradley, and B.N. Cuthbert, *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. Tech. Rep. No. A-6*. 2005, University of Florida: Gainesville.

[12] Lichtenauer, J., et al., *Cost-Effective Solution to Synchronized Audio-Visual Capture Using Multiple Sensors*, in *AVSS '09*. 2009, IEEE. p. 324-329.

[13] Morrison, G.S., et al., *Making Demonstrably Valid and Reliable Forensic Voice Comparison a Practical Everyday Reality in Australia - Database Collection Protocol*. 2010: ASSTA, Melbourne.