# AusTalk: an audio-visual corpus of Australian English

**Dominique Estival[1], Steve Cassidy[2], Felicity Cox[2], Denis Burnham[1]**
[1] MARCS Institute, U. of Western Sydney, Australia
[2] Macquarie University Australia
E-mail: d.estival@uws.edu.au, steve.cassidy@mq.edu.au, felicity.cox@mq.edu.au, d.burnham@uws.edu.au

**Abstract**

This paper describes the *AusTalk* corpus, which was designed and created through the *Big ASC*, a collaborative project with the two main goals of providing a standardised infrastructure for audio-visual recordings in Australia and of producing a large audio-visual corpus of Australian English, with 3 hours of AV recordings for 1000 speakers. We first present the overall project, then describe the corpus itself and its components, the strict data collection protocol with high levels of standardisation and automation, and the processes put in place for quality control. We also discuss the annotation phase of the project, along with its goals and challenges; a major contribution of the project has been to explore procedures for automating annotations and we present our solutions. We conclude with the current status of the corpus and with some examples of research already conducted with this new resource. *AusTalk* is one of the corpora included in the *Alveo* Virtual Lab, which is briefly sketched in the conclusion.

**Keywords:** audio-visual corpus, Australian English, standardised infrastructure, collection protocol, annotations.

## 1. The Big ASC project

The 'Big Australian Speech Corpus (Big ASC)' is a collaborative project between 11 institutions, funded by an Australian Research Council Linkage Infrastructure, Equipment and Facilities grant (Burnham, Cassidy, et al. 2009); (Burnham, Ambikairajah, et al. 2009) with the twin goals of 1) providing a standardized infrastructure for audio-visual (AV) recordings and 2) producing a large AV corpus of Australian English (AusE). Up to 1000 geographically and socially diverse speakers are recorded in locations across Australia using 12 sets of standardised hardware and software 'Black Boxes' with a uniform and automated protocol (the Standard Speech Collection Protocol – SSCP) to produce the *AusTalk* corpus (Wagner et al. 2010); (Burnham et al. 2011).

While the main purpose of *AusTal*k is to provide an extensible database for projects charting the extent, degree, and details of social, regional and ethno-cultural variations in AusE, it was also designed to support a range of research projects and applications. In order for the *AusTalk* corpus to cater to the needs of different researchers in various disciplines such as phonetics, forensic studies, speech and language technologies, linguistic analysis, audio-visual analysis, the Big ASC project had to strike a balance between high quality studio recording and field data collection. This was achieved through the strict data collection protocol, with high levels of standardisation and automation, and a recruitment process that ensured sufficient variability.

## 2. The *AusTalk* Corpus

When complete, the AusTalk corpus will comprise nearly 3000 hours of audio and video recordings from 1000 AusE speakers, all having completed primary and secondary education in Australia (but not necessarily having been born in Australia), a criterion that ensures inclusion of a range of speakers from various cultural backgrounds. According to the 2011 census, 26% of the Australian population were born overseas and a further 20% had at least one parent born overseas (ABS 2012). The criterion that participants had attended school in Australia was to ensure that we only captured speakers of AusE rather than foreign accented English. Demographic and language background information was collected from all participants via an online questionnaire to establish their gender, age, residential and educational history, cultural heritage, hobbies, occupation, languages spoken, parents' details and any speech and hearing difficulties they may have had. Almost 800 speakers have now been recorded at 13 different sites, with more than 2000 sessions uploaded, a total of 22TB of data. Data collection is still proceeding at three sites and will be completed by the end of 2014.

### 2.1 The *AusTalk* corpus components

Audio-visual corpora are important for different types of research in linguistics, Natural Language Processing (NLP) and Language and Speech Technologies, relying of various aspects of variability. In the *AusTalk* corpus, three one-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 locations covering all the capital cities of Australian states and territories and several regional centres. Stratified sampling across gender and three broad adult age groups captures *individual variability*. Wide advertising and high visibility of the project, with a well-publicised launch on Australia Day 2011 and good media coverage, helped recruit speakers from a range of social spheres to ensure *social variability*.

The *AusTalk* corpus contains a variety of speech content from a range of tasks, with four Read Speech and five Spontaneous Speech components, and all the data is captured by five microphones and two stereo cameras recording audio and video. Each of the three recording sessions comprises a different subset of the Read and Spontaneous speech tasks.

In the standard 'Words', 'Digits' and 'Sentences' tasks, the speaker reads aloud a list of prompts from a computer screen, while the 'Story Reading' and 'Story Re-telling' tasks (Session 1) provide material for the study

of differences between reading and spontaneous language. The 'Interview', 'Map Task' and 'Conversation' tasks provide material for the analysis of speech acts in dialogues. In the 'Interview' (Session 2), the speakers talk to the Recording Assistants (RAs) on a topic which they chose in Session 1. The 'Map Task' (Session 3) is designed along the lines of (Anderson et al. 1991) but adapted for AusE to contain locations and landmarks selected to sample a range of AusE phonological features. In this third session, two speakers are paired for two Map Tasks, so that each participant plays the role of Information Giver and Information Receiver, after which they discuss the experience in the 'Conversation'. At the beginning and end of each session, a set of natural 'Yes/No' questions elicit a range of positive and negative answers.

Table 1 shows the distribution of these tasks across the sessions and the average time for each task.

| Components | Session | Time (mins) | Time per speaker |
|---|---|---|---|
| **Read speech** | | | **53 mins** |
| Words (322 x 3) | S1, S2, S3 | 10 | 30 |
| Digit strings (12 x 2) | S1, S2 | 5 | 10 |
| Sentences (59 x 1) | S2 | 8 | 8 |
| Read story | S1 | 5 | 5 |
| **Spontaneous speech** | | | **80 mins** |
| Yes/No answers (x 5) | S1, S2, S3 | 2 | 10 |
| Re-told story | S1 | 10 | 10 |
| Interview | S2 | 15 | 15 |
| Map Task (x 2) | S3 | 20 | 40 |
| Conversation | S3 | 5 | 5 |
| **TOTAL (average)** | | | **133 mins** |

**Table 1: AusTalk Corpus Components / time per speaker**

## 2.2 Quality Control

To ensure high data quality as well as consistency across all the sites, several processes were put in place. First, before the data collection began, all the Recording Assistants (one for each of the 16 locations) were trained together during a 2-day centrally-located workshop at the University of Western Sydney (UWS), at which they practiced setting up the equipment and running through the recording sessions with each other. Additional training was required when new RAs were recruited, an important factor in maintaining consistency in the data collection.

Second, each recording site initially made sample recordings which were centrally checked for audio and video quality before the start of data collection at that particular site, and modifications were made to the location, to remove sources of noise or modify light conditions if the quality was sub-standard.

Third, there was continuous monitoring of data quality and feedback and advice to the RAs throughout the corpus collection. A Quality Control RA (QC-RA) was employed at the central UWS receiving site where the data was uploaded, with strict guidelines for both audio and video quality checks. To help the site RAs and the QC-RA, we developed the SSCP-QC, a utility to check the number of files along with the quality of parameters such as silence or loudness for audio, and frame skipping or brightness for video. The outcomes of the QC checks are retained and have become part of the published metadata indicating the QC status at the item and component levels. Manual inspection of the data finalises the published QC status, as one of the following:

- A (A-OK)
- B (OK, but imperfect)
- C (bad, not acceptable)
- D (deficient or missing, e.g. "`Missing 2nd video camera for Map Task`")

# 3. The AusTalk Annotation Task

Two important usability goals of the Big ASC project are to make *AusTalk* widely available and to allow future contributions, such as addition of further data or additional annotations. Audio and video data are stored on a web-accessible server, with corpus metadata and annotations stored in the DADA annotation store (Cassidy and Johnston 2009). The DADA server allows import/export of annotation data in formats supported by many annotation and analysis tools.

The Annotation Task itself could not be commenced until sufficient data were collected and organised (in late April 2012), but is now well under way. In this section, we first delimit the scope of the Annotation Task, then describe the processes we have put in place and the annotations that have already been produced before briefly mentioning the main challenges we faced.

The original goal of the Big ASC project was to annotate all the data collected and, from the beginning of the project, it was decided to automate the annotation process as much as possible, while providing high-quality manual annotation for a subset of the data. It was expected that forced alignment would be used where appropriate to enhance manual annotation. Thus, the Annotation Task

was limited to 1) word segmentation for the Read Speech and 2) orthographic transcription aligned at the phrase or sentence level for Spontaneous Speech. Integral to the project is that new annotations, e.g. detailed phonetic transcriptions or Part-of-Speech tagging, can later be contributed by project partners or other researchers and then integrated into the existing annotation store. We thus defined the goal for annotations as follows:

a) **Orthographic** level annotation of both Read and Spontaneous speech.
   For Read Speech, the script provides the basis for an orthographic transcription.
   For Spontaneous Speech, we were optimistic that orthographic transcripts could be generated from an automatic speech recognition package such a Dragon (DNS).
b) **Phonemic** level transcription with segmentation, to be automated as much as possible.
   For Read Speech, forced alignment would be performed in collaboration with an external partner (Schiel, Draxler, and Harrington 2011), following manual phonemic level transcription of a subset of the data. This subset would provide the training set for the forced aligner.
c) **Audio-Video** alignment. Automatic alignment is provided by the strobe signal recorded on a separate audio channel (Lichtenauer et al. 2009).

The following is the 'wish-list' of annotations to be performed if there were sufficient resources:

d) **Phonetic** level: manual and labour intensive
e) **Intonation**
f) **Part-of-Speech** – to be automated
g) **Morphemic**
h) **Syntactic**

We decided early on that the Big ASC project could only afford to manually annotate a subset of the data: for 5 speakers a full set of read speech data would be annotated (levels a and b above), while 100 additional speakers would only have a subset of their data manually annotated.

Manual annotation is labour-intensive and expensive. Therefore a major contribution of the project has been to explore procedures for automating annotations through collaboration with partners to produce 1) alignment for the Read Speech and 2) orthographic transcriptions for the Spontaneous Speech components.

As was expected, automating the time alignment of phonemic transcriptions for the Read Speech data proved very challenging. A major obstacle for this part of the project was that people make mistakes when reading material aloud, so scripted data does not always have the integrity required for automatic processing. The preliminary manual annotation phases of the project was therefore extended to provide sufficient high quality phonemically-transcribed data that would form the essential training materials for the Australian module of the Munich Automatic Segmentation System MAUS (Schiel, Draxler, and Harrington 2011). To this end the 59 read sentences from 100 speakers were orthographically and phonemically transcribed manually. The MAUS utility which was modified for AusE based on this training set is now capable of returning Praat textgrids (Boersma and Weenink 2001) containing phonemically transcribed and segmented data upon presentation of orthographic input. These automatically generated Praat textgrids can be manually corrected where necessary and provide a platform for more detailed segmentation and labelling in the future.

It proved too great a challenge to generate automatic orthographic transcriptions for Spontaneous Speech, so a third party transcription company was contracted to produce transcripts for the same sample of 100 speakers as above. It is then possible to pass the orthographic transcriptions through MAUS, which will return automatically generated textgrids for the Spontaneous Speech data.

In order to set a standard for annotation quality, the full set of scripted speech data from five speakers has been manually phonemically segmented and labelled in Praat by a team of highly trained annotators. We have created a purpose-built annotation manual to ensure consistency across annotators and tasks. Sentence data from an additional 30 speakers have also been manually annotated. These manually labelled data provide a benchmark that can be used as training material for further manual correction of automatically generated data.

## 4. Conclusion and Future Work

The data collection phase for *AusTalk* is coming to an end, with less than 200 speakers remaining to be recorded at three sites in order to complete the full complement of three one-hour sessions for 1000 AusE speakers.

Table 2 shows the distribution of speakers across recording sites. The figures given in the "Actual" columns show the number of speakers for whom data has been not only recorded but uploaded to the server. Detailed demographic statistics concerning gender, age and education level of the speakers are made available on the data server.

| STATE | Capital Cities (University) | Target | Actual | Regional Centres (University) | Target | Actual | Other | Target | Actual |
|---|---|---|---|---|---|---|---|---|---|
| **NSW** | Sydney (USYD) | 64 | **64** | Armidale (UNE) | 48 | **44** | Emotions (UNSW) | 36 | 0 |
| | Sydney (UNSW) | 48 | **48** | Bathurst (CSU) | 48 | **46** | | | |
| **QLD** | Brisbane (UQ) | 100 | **75** | Townsville (UQ) | 48 | **30** | | | |
| | | | | Maroochydore (USC) | 20 | **20** | | | |
| **VIC** | Melbourne (UMELB) | 120 | **118** | Castlemaine (UMELB) | 48 | **22** | | | |
| **SA** | Adelaide (Flinders) | 96 | **96** | | | | | | |
| **NT** | | | | Darwin (CDU) | 24 | **2** | Aboriginal English (CDU) | 48 | 0 |
| | | | | Alice Springs (CDU) | 24 | **0** | | | |
| **WA** | Perth (UWA) | 96 | **96** | | | | | | |
| **TAS** | Hobart (UTAS) | 48 | **48** | | | | | | |
| **ACT** | Canberra (UC) | 36 | **39** | | | | | | |
| | Canberra (ANU) | 48 | **48** | | | | | | |
| **Totals** | | 656 | **632** | | 260 | **164** | | 84 | 0 |

**Table 2. Distribution of recorded and uploaded speakers data (March 2014)**

Follow-on projects have already begun to collect data from different population groups in some locations (e.g. particular ethnic backgrounds in Canberra) and the analysis of *AusTalk* data is under way at other partner sites, e.g. video analysis for facial gestures (Sui et al. 2012a, 2012b) and close phonetic analysis of the isolated word list data.

There will be a tutorial at Interspeech 2014 (Togneri, Bennamoun, and Sui 2014) in which attendees will learn how to use the 3D based AV corpus derived from *AusTalk* for audio-visual speech/speaker recognition. Experimental results using this corpus show that, compared with the conventional AVSR based on the audio and grey-level visual features, there is a significant speech accuracy increase by integrating both depth-level and grey-level visual features.

In a study based on the framework of (Weiss, Burkhardt, and Geier 2013), the Read Sentences provide a rich body of stimuli used to study perceptual dimensions used by listeners to characterise speakers' vocal characteristics and speaking style. By collecting similarity measures for triplets of stimuli from 13 male speakers with an incomplete design (Burton and Nerlove 1976), fundamental perceptional dimensions separating these speakers can be extracted by applying multi-dimensional scaling on perception data from 15 male non-experts listeners. Along with the similarity decisions, individual labels for each triplet are assessed as a starting point for interpreting the dimensions found, as well as providing additional material to develop a questionnaire describing speakers' vocal characteristics and speaking style.

An unforeseen but very exciting addition to *AusTalk* was the inclusion of speakers who originally participated in the Australian National Database of Spoken Language (ANDOSL) project in 1993-95 (Vonwiller et al. 1995). Of the eight ANDOSL speakers who were found and who agreed to participate in AusTalk, four completed the full AusTalk recording sessions and these constitute invaluable longitudinal data for the study of AusE.

Annotation and quality assessment continue as more data are collected and made available through a new interface. Annotation is an important aspect of the Big ASC and other similar projects for, without it, many of the applications such as Automatic Speech Recognition, and much of the proposed research could not be conducted. While the ideal of providing full annotations of 100% of the data will not be realised in this phase of the project, we are providing a full set of manually created phonemic and orthographic transcriptions for a selected number of speakers. We will also provide automatically time-aligned transcriptions for all the Read Speech data and automatically generated orthographic transcriptions for at least a subset of the Spontaneous Speech data. Together these will constitute the basis and a protocol for further annotation in the future.

Meanwhile, the *AusTalk* corpus is already included in *Alveo,* the Human Communication Science Virtual Laboratory, a recent NeCTAR-funded Australian collaborative project that aims to provide a platform for easy access to language, speech and other communication-relevant databases and for the integrated use of a range of analysis tools (Burnham et al. 2012). *Alveo* incorporates existing tools, some developed by project members, which were adapted to work on the shared infrastructure, together with a web-based data discovery interface for searching and accessing the text,

speech, AV and music datasets contributed by the project partners. The tools are orchestrated by a workflow engine with both web and command line interfaces to allow use by technical and non-technical researchers (Cassidy et al. 2014). *Alveo* will allow the generation of automated Part-of-Speech tagging and syntactic analyses as additional annotations for the *AusTalk* corpus.

## 5. Acknowledgements

We gratefully 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.

## 6. References

ABS. 2012. Cultural Diversity in Australia. http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2071.0main+features902012-2013. Australian Bureau of Statistics.

Anderson, A.H., M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H.S. Thompson, and R. Weinert. 1991. "The HCRC Map Task Corpus." *Language and Speech* no. 34 (4):351-366.

Boersma, Paul, and David Weenink. 2001. "Praat, a system for doing phonetics by computer." *Glot International* no. 5 (9/10):341-345.

Burnham, Denis, E. Ambikairajah, Joanne. Arciuli, Mohammed Bennamoun, C.T. Best, Steven Bird, A.B. Butcher, Steve Cassidy, G. Chetty, F.M. Cox, Anne Cutler, Robert Dale, Julien R. Epps, Janet M. Fletcher, Roland Goecke, David B. Grayden, John T. Hajek, John C. Ingram, Shun Ishihara, Nenagh Kemp, Yuko Kinoshita, T. Kuratate, T.W. Lewis, D.E. Loakes, Mark Onslow, David M. Powers, P. Rose, Roberto Togneri, D. Tran, and Michael Wagner. 2009. A blueprint for a comprehensive Australian English auditory-visual speech corpus. In *2008 HCSNet Workshop on Designing the Australian National Corpus*. Sydney: Somerville, MA, USA: Cascadilla Proceedings Project.

Burnham, Denis, Steve Cassidy, Felicity Cox, and Robert Dale. 2009. The Big Australian Speech Corpus: An Audio-Visual Speech Corpus of Australian English Australian Re-search Council Linkage, Infrastructure, Equipment and Facilities Grant. Original edition, LE100100211.

Burnham, Denis, Dominique Estival, Peter Bugeia, Peter Sefton, and Steven Cassidy. 2012. Above and Beyond Speech, Language and Music: A Virtual Lab for Human Communication Science (HCS vLab). NeCTAR (National eResearch Collaboration Tools & Resources) Virtual Laboratory. Original edition, VL222.

Burnham, Denis, Dominique Estival, Steven Fazio, Felicity Cox, Robert Dale, Jette Viethen, Steve Cassidy, Julien Epps, Roberto Togneri, Yuko Kinoshita, Roland Göcke, Joanne Arciuli, Marc Onslow, Trent Lewis, Andy Butcher, John Hajek, and Michael Wagner. 2011. Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box. In *Interspeech 2011*. Florence, Italy.

Cassidy, Steve, Dominique Estival, Timothy Jones, Denis Burnham, and Jared Burghold. 2014. The Human Communication Science Virtual Laboratory: A Web Based Repository API. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, Iceland.

Cassidy, Steve, and Trevor Johnston. 2009. Ingesting the Auslan Corpus into the DADA Annotation Store. In *Third Linguistic Annotation Workshop (LAW III)*. Singapore.

Lichtenauer, Jeroen, Michel Valstar, Jie Shen, and Maja Pantic. 2009. Cost-Effective Solution to Synchronized Audio-Visual Capture Using Multiple Sensors. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09)*. Washington, DC, USA: IEEE Computer Society.

Schiel, Florian, Christoph Draxler, and Jonathan Harrington. 2011. Phonemic Segmentation and Labelling using the MAUS Technique. In *Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research'*. University of Pennsylvania, Philadelphia PA. USA.

Sui, Chao, Serajul Haque, Roberto Togneri, and Mohammed Bennamoun. 2012a. A 3D Audio-Visual Corpus for Speech Recognition. In *SST2012*. Sydney, Australia: ASSTA.

Sui, Chao, Serajul Haque, Roberto Togneri, and Mohammed Bennamoun. 2012b. Discrimination Comparison Between Audio and Visual Features. In *Asilomar 2012*. Pacific Grove, USA.

Togneri, Roberto, Mohammed Bennamoun, and Chao Sui. 2014. Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus. Tutorial at Interspeech 2014. Singapore.

Vonwiller, J., I. Rogers, C. Cleirigh, and W. Lewis. 1995. "Speaker and Material Selection for the Australian National Database of Spoken Language." *Journal of Quantitative Linguistics* no. 3:177-211.

Wagner, M., D. Tran, R. Togneri, P. Rose, D. Powers, M. Onslow, D. Loakes, T. Lewis, T. Kuratate, Y. Kinoshita, N. Kemp, S. Ishihara, J. Ingram, J. Hajek, D.B. Grayden, R. Göcke, J. Fletcher, D. Estival, J. Epps, R. Dale, A. Cutler, F. Cox, G. Chetty, S. Cassidy, A. Butcher, D. Burnham, S. Bird, C. Best, M. Bennamoun, J. Arciuli, and E. Ambikairajah. 2010. The Big Australian Speech Corpus (The Big ASC). In *13th Australasian International Conference on Speech Science and Technology*, edited by M. Tabain, J. Fletcher, D. Grayden, Hajek J. and A. Butcher. Melbourne: ASSTA.