A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus

Denis Burnham[a1], Eliathamby Ambikairajah[b], Joanne Arciuli[c], Mohammed Bennamoun[d], Catherine T. Best[a], Steven Bird[e], Andrew R. Butcher[f], Steve Cassidy[g], Girija Chetty[h], Felicity M. Cox[g], Anne Cutler[a,i], Robert Dale[g], Julien R. Epps[b], Janet M. Fletcher[e], Roland Goecke[h,j], David B. Grayden[e], John T. Hajek[e], John C. Ingram[k], Shunichi Ishihara[j], Nenagh Kemp[l], Yuko Kinoshita[h], Takaaki Kuratate[a], Trent W. Lewis[f], Debbie E. Loakes[e], Mark Onslow[c], David M. Powers[f], Philip Rose[j], Roberto Togneri[d], Dat Tran[h], and Michael Wagner[h]

[a]University of Western Sydney, [b]University of New South Wales, [c]University of Sydney, [d]University of Western Australia, [e]University of Melbourne, [f]Flinders University, [g]Macquarie University, [h]University of Canberra, [i]Max Planck Institute for Psycholinguistics (Nijmegen), [j]Australian National University, [k]University of Queensland, [l]University of Tasmania

1. Introduction and Rationale

Contemporary speech science is driven by the availability of large, diverse speech corpora. Such infrastructure underpins research and technological advances in various practical, socially beneficial and economically fruitful endeavours, from ASR to hearing prostheses. Unfortunately, speech corpora are not easy to come by because they are both expensive to collect and are not favoured by the usual funding sources as their collection *per se* does not fall under the classification of 'research'. Nevertheless they provide the *sine qua non* for many avenues of research endeavour in speech science.

The only publicly available Australian speech corpus is the 12-year-old Australian National Database of Spoken Language (ANDOSL) database (see http://andosl.anu.edu.au/; Millar, Dermody, Harrington, & Vonwillar, 1990), which is now outmoded due to its small number of participants, just a single recording session per speaker, low fidelity, audio-only rather than AV data, its lack of disordered speech, and limited coverage of indigenous and ethnocultural Australian English (AusE) variants. There are more up-to-date UK and US English language corpora, but these are mostly audio-only, and use of these for AusE purposes is not optimal, and results in inaccuracies.

2. Purpose of the Big Australian Speech Corpus (The Big ASC)

In Australia we have significant research strengths in speech science that require an extensive AusE AV speech corpus. However, currently there is none. Here we describe a blueprint for establishing the Big Australian Speech Corpus (the Big ASC), a corpus of over 1,100 speakers from all over Australia. With the support of the Human Communication Science Network and the Australasian Speech Science and Technology Association, speech science experts from across Australia have banded together to plan the recording of large quantities of AV speech from many locations and

[1] For enquiries, please contact Prof Denis Burnham, MARCS Auditory Laboratories, University of Western Sydney, d.burnham@uws.edu.au

multiple sessions using (a) standard recording equipment, (b) a standard collaboratively designed protocol, and (c) storage and annotation in an existing/developing Distributed Access and Data Annotation system. With a projected lifespan of at least two decades, the Big ASC would engender and enhance Australian research in a range of human communication and speech science areas. A representative selection of these areas is set out below.

2.1. Phonetics and Linguistics

The Big ASC is essential to describe the variation of AusE over geographical area (Butcher, 2006, 2008; Cox & Palethorpe, 1998, 2001, 2004), ethnocultural and social background, and speech style (Ingram, 1989); to describe changes to the language since the collection of the outmoded ANDOSL database; and to provide greater access to information on speech production (Fletcher, Grabe, & Warren, 2004).

2.2. Psycholinguistics

The Big ASC would have applications in projects on psycholinguistic models for word processing (Cutler & Carter, 1987; Cutler, 2005); young children's perception of phonetic variability and dialectal variation in spoken words (Best, Tyler, Gooding, Orlando, & Quann, in press); the effect of pronunciation on written language (Kemp, 2009); and hearing training programs for children and adult users of cochlear implants (Dawson, McKay, Busby, Grayden, & Clarke, 2000; Mok, Grayden, Dowell, & Lawrence, 2006).

2.3. Engineering – Spoken Language Processing

The corpus would support research projects in ASR and AV ASR (Lewis & Powers, 2005, 2008; Saragih & Goecke, 2007); the Thinking Head project (see Section 4.3 and http://thinkinghead.edu.au/); speaker authentication and localisation based on a fusion (Lewis & Powers, 2005, 2008) and separation (Li & Powers, 2001) of multiple signals including voice acoustics and facial image in particular (Tran, Wagner, Lau, & Gen, 2004; Tran & Wagner, 2002); automatic real-time visual biometric systems robust to variations; development of more robust systems for authentication or identification (e.g., government and commercial services such as Centrelink and telephone banking) available in 4G mobile telephony (Naseem, Togneri, & Bennamoun, 2009); cochlear implant sound processing for improved perception of speech in noise and access to speaker identity and intonation (Bavin, Grayden, Scott & Stefanakis, in press; Talaricoa et al., 2007); and emotion detection applications, (e.g., determining 'choice points' for automatic user service systems switching to a manual operator, or Talking Heads switching between language and dialog models) (McIntyre & Goecke, 2007; Yacoub, Simsky, Lin, & Burns, 2003; Vidhyasaharan, Ambikairajah, & Epps, 2009); and AV TTS synthesis (Kuratate, 2008).

2.4. Language Technology and Computer Science

In this area, various interfaces would be enabled, for example, ASR tailored for AusE and its variety of accents and emotional tones/textures/expressions (Powers et al., 2008), speech dialogue management (Dale & Viethen, 2009; Viethen & Dale, 2006) and AV user-centric/context-aware/ask-once/ask-nonce information retrieval and monitoring (Powers & Leibbrandt, 2009); as well as web search and training products and guides based on grounded speech understanding (Huang & Powers, 2008; Pfitzner, Leibbrandt, & Powers, 2008; Pfitzner, Treharne, & Powers, 2008).

2.5. Speech Pathology

Corpora of disordered speech *and* representative Australian speech are critical to describe and analyze disordered speech, understand the disorders, and develop intervention treatments and devices (Butcher, 1996; Arciuli & McLeod, 2008).

2.6. Forensic Speech Science

Spontaneous speech from multiple sessions would allow estimation of between- and within-speaker variability across different recording sessions. This allows estimation of the strength of evidence with a likelihood ratio using Bayes theorem (Rose, 2002). The Big ASC would be of great use in testing forensic speaker recognition approaches and conducting real-world casework, as well as identifying individuality in speaker behaviour (Butcher, 2002; Loakes & McDougall, in press).

3. Design of the Big ASC: An Overview

Input from Australian experts who would be using the Big ASC is crucial for the construction of a comprehensive, maximally applicable corpus. To date, 29 speech scientists from 11 Australian universities have contributed their disciplinary expertise to devise optimal equipment and protocols. The Big ASC infrastructure would provide a significant boost to speech research in Australia now and well into the future because it would incorporate contemporary and rigorous design features as follows.

3.1. Tight Control

Standardisation in equipment and data collection procedures is essential. A Standard Speech Science Infrastructure Black Box (SSSIBB) and a Standard Speech Collection Protocol (SSCP) would be used at each collection node to ensure that speech collection conditions are controlled and documented.

3.2. High-Fidelity

Two-channel AV recording would allow spatial localisation, and both auditory scene analysis and 3D imaging.

3.3. Size and Distribution

Large speech corpora (e.g., Smits, Warner, McQueen, & Cutler, 2003) are essential in order to cover idiosyncrasies and variation. Here speech from over 1,100 speakers from 11 collection nodes from every state and territory of Australia would be collected.

3.4. Multiple Sessions (Within-Speaker Variation)

Each speaker would be recorded on three separate occasions (Σ=3384 sessions) to capture within-speaker variability over time.

3.5. Diversity (Between-Speaker Variation)

Representative sampling from 11 different nodes would reflect *regional* (all states and territories), *indigenous* (varieties of Aboriginal English and 2 creoles), and *ethnocultural* (AusE from Greek, Italian, Lebanese, and Chinese background speakers) variation, and degree of *intactness* (disordered speech).

3.6. AV data

The increased power of modern computers, the overwhelming evidence of efficacy of visual speech information in disambiguating speech and speaker recognition (Benoît, Lallouache, Mohamadi, & Abry, 1992; Girin, Feng, & Schwartz, 2001; Potamianos, Neti, Luettin, & Matthews, 2004), and the currency and topicality of AV avatars and embodied conversational agents in Talking Heads mean that it is now *de rigueur* for speech corpora to be AV (note, for example, the Advancing Video Audio Technology (AVAtech) project at the Max Planck Institute for Psycholinguistics; AVAtech, 2009).

3.7. Efficient Management

The Big ASC would use and extend an existing/developing language data storage system, the Distributed Access and Data Annotation for the Human Communication Sciences (DADA-HCS) (see Section 4.4), to provide shared access to the corpus and the collective annotation and other metadata associated with every recording.

3.8. Australian

This would be the first Australian speech corpus to meet the demands of modern speech science and would sample widely and appropriately from the breadth of AusE variations.

4. Support for the Big ASC

The Big ASC blueprint builds on, is supported by, and will support relevant associations, networks, and projects as set out below.

4.1. Australasian Speech Science and Technology Association (ASSTA)

ASSTA advances the understanding of speech science and technology both within Australia (e.g., biennial Speech Science and Technology (SST) conference and a range of research funding initiatives) and internationally via interaction with the International Speech Communication Association (ISCA). Within ASSTA, two subcommittees would provide leadership and specialist knowledge: the National Spoken Language Database (NSLD) subcommittee in the main, as well as the Forensic Speech Science subcommittee (FSSC) where forensic matters are concerned.

4.2. Human Communication Science Network (HCSNet)

HCSNet is an Australian Research Council (ARC) research network jointly run by the University of Western Sydney and Macquarie University. HCSNet brings together a wide mix of researchers who work on speech, text, and sonics, including those working on the Big ASC project. In addition to this corpus project, HCSNet has spawned other large projects such as the DADA-HCS and the Thinking Head project.

4.3. ARC/NHMRC Special Initiatives Thinking Systems Project 'From Talking Heads to Thinking Heads'

This project brings together human communication scientists from six Australian and three international universities, and integrates best-practice talking-head science and technology with behavioural evaluation and performance art to provide a plug-and-play Thinking Head research platform. Within this, speech science applications relying on speech corpora (ASR, TTS synthesis, dialog, animation) can be compatibility tested and evaluated for user satisfaction and engagement.

4.4. Distributed Access and Data Annotation for the Human Communication Sciences (DADA-HCS)

DADA-HCS was spawned by HCSNet and has been adopted by the Thinking Head project for data management. It would also be used here for data storage, annotation, and access (see Section 5.5).

5. Main Components of the Big ASC

5.1. Sampling Variation

For a good speech corpus with wide applicability a surfeit of speech variation is mandatory (Smits et al., 2003). The Big ASC would incorporate a wide range of speakers and locations (see Table 1 for possible data collection sites and sampling breakdown).

Table 1.
Possible Data Collection Sites and Roles Involved in Establishing the Big ASC

Site	**Data Collected *or Role***
Hobart	Standard AusE (n=72) Regional AusE (n=24)
Perth	Standard AusE (n=96)
Adelaide (1)	Standard AusE (n=96)
Adelaide (2)	Aus Indigenous Eng. (n=48) AusE-Indig. Creoles (n=48)
Melbourne	Standard AusE (n=96) Italian AusE (n=48) Greek AusE (n=48)
Canberra	Standard AusE (n=96) Standard AusE (n=96)
Brisbane	Standard AusE (n=96) Regional AusE (n=24)
Sydney (1)	Standard AusE (n=48) Chinese AusE (n=48)
Sydney (2)	Disordered AusE (n=96)
Sydney (3)	Standard AusE (n=48) Regional AusE (n=48)
Sydney (4)	*DADA Implementation & Annotation HQ*
Sydney (5)	Lebanese AusE (n=48) *Project Administration*

The rationale for the variation in sampling participants and procedures for obtaining representatives samples are detailed below:

5.1.1. Regional and Ethnocultural Variation

A representative sample of adult male and female speakers of non-indigenous AusE across the country in three age groups (<25, 30-45, >50) and two socioeconomic levels would be collected. In Adelaide, Sydney, Perth, Brisbane, Melbourne, Hobart (including some regional areas surrounding Hobart) and Canberra, 16 speakers (8 females, 8 males) would represent the six age x socioeconomic combinations (n=96), a total of N=672 speakers. In each of two regional areas in NSW, and in Townsville, data would be collected from four males and females from the three age groups (n=24, N=72). (Data from additional regional areas across Australia would be collected at a later stage, further funds permitting.) Finally, the four largest ethnocultural groups of Australian-born citizens with parents from non-English speaking countries - Italian (11%), Greek (6%), Chinese (6%), and Lebanese (3%) (Australian Bureau of Statistics 2006 census) - would be sampled in Sydney (Chinese and Lebanese) and Melbourne (Italian and Greek) from males and females in three age groups (n=48, N=192). This is a total of 744 speakers incorporating regional variations of Standard AusE and 192 with ethnocultural variations.

5.1.2. Aboriginal English Variation

The majority of Australia's 455,000-strong Aboriginal population speak some form of Australian Aboriginal English (AAE) and it is the first (and only) language of a large number of Aboriginal children. Thus their language is somewhere on a continuum from something very close to Standard AusE through to creole. There are two distinct creoles – one spoken in the Torres Strait (TS) Islands and TS Islander communities in Queensland (23,000 speakers), and the other, 'Kriol', on the mainland from the Kimberley through the Barkly Tableland to the Queensland gulf country (20,000 speakers; National Indigenous Languages Survey report, 2005). Like all other creoles, these are languages in their own right with complex, rule-governed codes and extensive vocabulary. Recordings would be made in Darwin, Alice Springs, Fitzroy Crossing (12 AAE & 12 Kriol speakers, 6 male, 6 female) and on Waibene (Thursday Island) (12 AAE and 12 TS Creole speakers, 6 male, 6 female).

5.1.3. Disordered Speech Variation

In the USA occupations are voice-dependent for 34% of workers (87.5% of workers in large urban areas) and the economic cost of communication disorders is $154.3-186B per anum (Ruben, 2000). There are no equivalent data for adults in Australia, but a recent study of 14,500 Australian primary and secondary school students suggests that the prevalence of communication disorders is around 13% (McLeod & McKinnon, 2007). One particularly common speech disorder is stuttering, which develops unpredictably and rapidly during early childhood, disturbs peer interactions (Langevin, Packman, Thompson, & Onslow, 2009), and can be associated with occupational under-achievement, impaired oral communication, and a high level of social phobia (Australian Stuttering Research Centre cohort; Menzies et al., 2008). Speech data from 96 stutterers would be collected representatively, if possible, from the three age groups and two socioeconomic levels, and the greater incidence of stuttering in males than females may be reflected in the final sample.

5.1.4. Participants and Field Trips

A total of 1,100 participants would be required with an approximate budget of around $1000 per recording node for advertising/recruitment and reimbursement of travel expenses (with three visits per participant). Field trips would be essential for the required diversity of Big ASC, and probable locations would include (a) Adelaide to Darwin, Alice Springs, Fitzroy Crossing, and Waibene (Thursday Island) for the collection of Australian Aboriginal English data and AusE-indigenous creoles; (b) Sydney to Broken Hill and Longreach for regional AusE; (c) Brisbane to Townsville for regional AusE; and (d) Tasmania for regional AusE data collection.

5.2. Standard Speech Science Infrastructure Black Box (SSSIBB)

Standardization is also necessary with regard to equipment; an SSSIBB would be established at each participating recording site. This integrated piece of hardware would be comprised of a portable computer, stereo cameras and stereo microphones, and a 360° camera to ensure compatibility of audio and video data streams between recording sites and a record of the wider recording context.

5.3. Standard Speech Collection Protocol (SSCP)

A variety of tasks appropriate for different applications would be completed across three separate recording sessions (see Table 2.) As literacy in English (or creole) cannot be assumed for the Australian Indigenous sample, some variation of the protocol would be necessary: Sentences and word lists would be orally prompted, the Map tasks and transcript readings replaced by alternative tasks such as story telling, and the Emotional Speech task could be modified or omitted. Importantly, all the word-level and natural sentence-level material would be retained. The rationale for particular components of the SSCP is set out below.

5.3.1. Phonetic and Style Variation

Comprehensive demographic, family, and historical data would be collected in the first session to document the regional and ethnocultural dialect variations of each speaker. Participants would be recorded on three separate occasions to allow natural variation in voice quality in a range of speech situations. The time between sessions would be short (1 week between sessions 1 and 2), and longer (4 weeks between 2 and 3) (some reductions could be required on field trips). Core data collection tasks would elicit formal speech and contain standard digit and word lists (the HvD task; 'h'-vowel-'d' words, e.g., 'had', 'hid', etc.) and phonetically balanced Read Sentences material, the latter both in natural and emotional speech. Noncore data collection would capture unguarded dialogue, conversational speech, and style shifting. A particularly good indicator of style shifting would be the spontaneous narrative in Session 2 (elicited after the Interview task by a request to relate a particularly dangerous or exciting anecdote or experience) versus a version of the same text in Session 3 spoken in 'newsreader' style from a transcript of the narrative made by a research assistant (RA) between the second and the third sessions.

5.3.2. Forensic Speaker Recognition (FSR)

The yes/no elicitation item would provide natural variations for 'yes' ("yes, yeah, yep"), 'no' ("no, nah"), 'um' ("ah, mm"), words very useful in forensic casework (Rose, 2002; Arciuli, Mallard, & Villar, in press). The Map task involves two people, visually shielded from each other, each having access to a map which has some information common to both maps and some peculiar to each with one participant guiding the other to a particular destination. (Only one participant would be recorded audio-visually using the standard SSSIBB apparatus (see Section 5.2), while the other would be just audio-recorded. The Map task would be conducted at the end of a session for participant A, and the start of one for participant B, and would be repeated in sessions 1 and 2, so A and B can be the subject of AV or just audio recording in each session.) Incorporated into the task are long, difficult place names with participants being asked to spell these, and fictional addresses and names to elicit speech segments in a spontaneous yet controlled fashion. Telephone speech is important for forensic applications. Telephones severely attenuate low frequencies of speech, including the fundamental, so pitch must be perceived via upper harmonics - the 'missing fundamental effect'. They also severely attenuate high frequency components, which contain speaker-specific information, for example in third and higher formants. Telephone speech would be obtained by passing Read Sentences task speech through various filters (codecs for regional and commercial variations of mobile phones, landlines).

5.3.3. Speech/Speaker Recognition

In the Read Sentences task, varied consonant/vowel coarticulation combinations are important for extraction of diphones for acoustic models in ASR, as is the repeated HvD task, and the Digits task is important for speaker verification in voice password situations. The Map task, and the Interview and Spontaneous Narrative (in which the RA would ask open questions to allow spontaneous speech, then segue to the elicitation of a spontaneous narrative) are essential for collecting connected spontaneous speech for constructing prosody models and setting up language models for ASR and dialog management. The Speech-in-Noise task involves the participant speaking through multispeaker babble, resulting in hyperarticulated speech. The Read Sentences task would be used for comparison with clear speech. Speech-in-noise data are particularly useful for training ASR and systems in suboptimal (real world) conditions.

Table 2.
Standard Speech Collection Protocol (SSCP) for Sessions at all Recording Nodes

	Session			Annotation
	1st	**2nd (1 week later)**	**3rd (4 weeks later)**	
Initial	Demographic, consent, ethno-cultural questionnaire			
Core Material	Calibration (sound & light, time readings)	Calibration (sound & light, time readings)	Calibration (sound & light, time readings)	
	AV speech calibration	AV speech calibration	AV speech calibration	
	Digits	Digits	Digits	Word
	HvDs (+laterals & nasals)	HvDs (+laterals & nasals)	HvDs (+laterals & nasals)	Vowel
	Read Sentences	Read Sentences	Read Sentences	Phoneme
	Emotion Sentences	Emotion Sentences	Emotion Sentences	Phoneme
	Yes/No elicitation	Yes/No elicitation	Yes/No elicitation	Word
x 1 Extra Material			Speech-in-noise	Word
		Interview		Turns
x 2 Extra Material		Spontaneous narrative	Reading transcript of previous narrative	Transcript
	Map task #1	Map task #2		Transcript

5.3.4. Emotional Speech

As an extension of Read Sentences, participants would be requested to read a given sentence according to one of seven emotions (neutral, anger, happiness, sadness, fear, boredom, and stressed). Then, as a variation of the Interview task, participants would be asked to converse naturally with the RA in each of the seven emotions (as in the Read Sentences task). The naturalness of elicited emotion based on this methodology is inherently limited. Historically, much emotional speech data has been produced by actors (Douglas-Cowie, Campbell, Cowie, & Roach, 2003), with the advantage that human recognition performance benchmarks for such data are high. The proposed approach foregoes some naturalness and typicality of actual language use, in favour of greater certainty with respect to the emotion produced than would be possible from a less structured task, and in favour of control over the linguistic content (in the case of Read Sentences). A critical deficiency in the currently available emotional speech corpora is the lack of accurate phonetic labeling. Given the time required for the Emotion Interview task, it would be conducted only at one Sydney site with the 48 speakers of standard AusE to be tested there. In many cases, time (1-2 minutes) would be required for participants to practice producing a given emotion, and this protocol has been used in previous less extensive studies (LDC Emotional Prosody Speech Corpus, 1992).

5.3.5. AV Speech

AV speech data are essential for many applications, for example, ASR and speaker recognition, biometric password applications. All data (except ½ of the Map task on each occasion) would be AV-recorded, and the initial lateral head movements (AV calibration) would facilitate recording AV speech. The Speech-in-Noise and the Emotion tasks are of particular interest for mapping between auditory and visual components of hyperarticulated speech and emotional speech respectively, and the development of smarter ASR and talking heads.

5.4. Annotation

A base level of annotation of data would be conducted by node RAs at each site. For recordings that are read (Digits, Read Sentences etc.), the start and end of each word would be marked, while for the longer unscripted recordings, a transcript of what is said would be aligned at the phrase or sentence level. In addition, the node RAs would transcribe the participants' spontaneous narrative in Session 2 to allow a 'newsreader' version of the same text in Session 3. Validation of the basic node level annotation, together with more detailed annotation, would be conducted by the central annotation team. Consistent principles and protocols for annotation would be determined. The annotation team would, for example, mark up aspects of dialogue, intonational, syntactic, and rhetorical structure as appropriate. Annotation would involve variants of the Emu and ELAN tools, which would be interfaced with the shared annotation server running the DADA-HCS system to be used by all annotators in the project in order to build the corpus collaboratively and consistently.

5.5. Distributed Access and Data Annotation for the Human Communication Sciences (DADA-HCS)

The DADA-HCS project has developed a distributed data store designed to make shared access to large collections of language data easier. DADA-HCS (ARC Grant SR0567319; Cassidy & Ballantine, 2007; Cassidy, 2008) allows data to be shared efficiently among project members and manages shared access to the annotated data so that multiple parties can develop a definitive annotation collaboratively. The Big ASC would support and be supported by the DADA-HCS system, using and extending it to provide shared access to the corpus. The Big ASC would form not only an intact piece of infrastructure, but one that is embedded in the DADA-HCS system that affords future augmentation by the project investigators (using the hardware used and protocols established in this project) and by others, so that further subsamples, for example, child speech, may be included later.

5.6. Servers and Back-up

A Central/Primary Data Store/Server is essential to hold the very large amounts of AV data. It would have the appropriate RAID disk storage devices and media, and would also be used for software development and quality control. A back-up Secondary Data Store/Server would also have the appropriate RAID disk storage devices and media, and be used for e-annotation.

5.7. Personnel

A project manager/software engineer would be essential to coordinate corpus collection; direct and support its annotation and subsequent dissemination; oversee the technical coordination of the project; and provide assistance to individual sites where needed. A programmer would be required to build software for the data collection, including AV recording and entry of metadata for each recording session. The programmer would extend the DADA-HCS system to support collaborative annotation of the data, and integrate the DADA-HCS back-end with the Emu Speech Database System (Cassidy, 1998) to provide annotation tools for the project. The programmer would be responsible for the central data store and support the collaborative annotation of the data. RAs would be required for general

administration, running recording sessions, constructing the first level of metadata, and conducting some transcription and labelling of the recorded data. For the more difficult continuous speech samples, a small band of annotation specialist RAs would be required. At each node there would be a chief investigator for overseeing the project, coordinating hardware and space issues for testing, and supervising the RA.

6. Funding the Big ASC

The total cost of building the Big ASC is estimated to be in excess of $1.5 million. A possible funding source is the ARC Linkage Infrastructure and Equipment (LIEF) scheme, and an application will be submitted for this scheme in 2009, requesting 75% of the project costs with the other 25% coming from participant universities.

7. Acknowledgements

The authors would like to thank Amanda Reid for assistance with manuscript preparation and editing, and coordinating author contributions.

References

Arciuli, Joanne, David Mallard, & Gina Villar (In press) "Um, I can tell you're lying": Linguistic markers of deception vs. truth-telling in speech. *Applied Psycholinguistics.* Accepted May 2009.

Arciuli, Joanne, & Sharynne McLeod. (2008). Production of /st/ clusters in trochaic and iambic contexts by typically developing children. In Rudolph Sock, Susanne Fuchs, & Yves Laprie (Eds.), *Proceedings of the 8th International Seminar on Speech Production (ISSP)* (pp. 181-184)*.* Strasbourg, France: INRIA.

Advancing video audio technology (AVAtech) project. (2009). Max Planck Institute for Psycholinguistics. Available at: http://www.mpi.nl/research/research-projects/language-archiving-technology/news/avatech-advancing-video-audio-technology-in-humanities-research-project

Bavin, Edith L, David B. Grayden, Kim Scott, & Toni Stefanakis. (In press). Testing auditory processing skills and their associations with language in 4-5 year-olds. *Language & Speech*, *53*. Accepted October 2008.

Benoît, Christian, Tahar Lallouache, Tayeb Mohamadi, & Christian Abry. (1992). In Gerard Bailly & Christian Benoît (Eds.), *Talking machines* (pp. 485-504). Amsterdam: North Holland.

Best, Catherine T., Michael D. Tyler, Tiffany N. Gooding, Corey B. Orlando, & Chelsea A. Quann. (In press). Emergent phonology: Toddlers' perception of words spoken in non-native vs native dialects. *Psychological Science*.

Butcher, Andrew R. (1996). Levels of representation in the acquisition of phonology: Evidence from 'before and after' speech. In Barbara Dodd, Ruth Campbell, & Linda Worall (Eds.), *Evaluating theories of language: Evidence from disordered communication* (pp. 55-73). London: Whurr Publishers.

Butcher, Andrew R. (2002, July). Forensic phonetics: Issues in speaker identification evidence*. Proceedings of the Inaugural International Conference of the Institute of Forensic Studies: Forensic Evidence: Proof and Presentation.* Prato, Italy [CD-ROM no page numbers].

Butcher, Andrew R. (2006). Formant frequencies of /hVd/ vowels in the speech of South Australian females. In Paul Warren & Catherine I. Watson (Eds.), *Proceedings of the 11th Australasian International Conference on Speech Science & Technology* (pp. 449-453). Auckland, New Zealand: Australian Speech Science & Technology Association Inc.

Butcher, Andrew R. (2008). Linguistic aspects of Australian Aboriginal English. *Clinical Linguistics & Phonetics*, *22*, 625-642.

Cassidy, Steve. (1998). Emu Speech Database System V 1.2, available at http://emu.sourceforge.net/manual/manual.html/

Cassidy, Steve, & James Ballantine. (2007, July). *Version control for RDF triple stores.* Paper presented at the 2nd International Conference on Software and Data Technologies, Barcelona.

Cassidy, Steve. (2008, May). *A RESTful interface to annotations on the web.* Paper presented at the *2nd Linguistic Annotation Workshop*, Morocco.

Cox, Felicity, & Sallyanne Palethorpe. (1998). Regional variation in the vowels of female adolescents from Sydney. In Robert H. Mannell & Jordi Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing.* Sydney: Australian Speech Science and Technology Association, Incorporated.

Cox, Felicity, & Sallyanne Palethorpe. (2001). The changing face of Australian English vowels. In D. Blair & P. Collins (Eds.), *Varieties of English around the world: English in Australia* (pp. 17-44). Amsterdam: John Benjamins Publishing.

Cox, Felicity, & Sallyanne Palethorpe. (2004). The border effect: Vowel differences across the NSW/Victorian border. In Christo Moskovsky (Ed.), *Proceedings of the 2003 Conference of the Australian Linguistic Society* (pp. 1-14). Australian Linguistic Society. http://www.als.asn.au.

Cutler, Anne, & David Carter. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language, 2,* 133-142.

Cutler, Anne. (2005). The lexical statistics of word recognition problems caused by L2 phonetic confusion. *Proceedings of Interspeech 2005 – Eurospeech 2005, 9th European Conference on Speech Communication and Technology* (pp. 413-416)**.** Lisbon, Portugal: International Speech Communication Association. Available at http://www.isca-speech.org/archive/interspeech_2005

Dale, Robert, & Jette Viethen. (2009). Referring expression generation through attribute-based heuristics. In Emiel Krahmer & Mariet Theune (Eds.), *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 58-65). Stroudsburg, PA, USA: The Association for Computational Linguistics.

Dawson, Pam W., Colette M. McKay, Peter A. Busby, David B. Grayden, & Graeme M. Clark. (2000). Electrode discrimination and speech perception in young children using cochlear implants. *Ear and Hearing, 21,* 597-607.

Douglas-Cowie, Ellen, Nick Campbell, Roddy Cowie, & Peter Roach (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40*(3), 33-60.

Fletcher, Janet, Esther Grabe, & Paul Warren. (2004). Intonational variation in four dialects of English: The high rising tune. In Sun-Ah Jun (Ed.), *Prosodic typology* (pp. 390-409). Oxford: Oxford University Press.

Girin, Laurent, Gang Feng, & Jean-Luc Schwartz. (2001). Audiovisual enhancement of speech in noise. *Journal of the Acoustical Society of America, 109,* 3007-3020.

Huang, Jin Hu, & David M. W. Powers. (2008). Suffix-tree-based approach for Chinese information retrieval. *Proceedings of the International Conference on Intelligent Systems Design and Applications, vol. 3* (pp. 393-397). Washington, DC, USA: Institute of Electrical and Electronics Engineers Computer Society.

Ingram, John. (1989). Connected speech processes in Australian English. In David Bradley, Roland D. Sussex, & Graham K. Scott (Eds.), *Studies in Australian English* (pp. 21-49). Bundoora, Victoria : Dept. of Linguistics, La Trobe University for the Australian Linguistic Society.

Kemp, Nenagh. (2009). The spelling of vowels is influenced by Australian and British English dialect differences. *Scientific Studies of Reading, 13,* 53-72.

Kuratate, Takaaki. (2008). Text-to-AV synthesis system for thinking head project. Pp 191-194. In Roland Goecke, Patrick Lucey, & Simon Lucey (Eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008* (pp. 191-194). Moreton Island, Australia: Auditory-Visual Speech Association.

LDC Emotional Prosody Speech Corpus. (1992). Linguistic Data Consortium, University of Pennsylvania.

Langevin, Marilyn, Ann Packman, Robyn Thompson, & Mark Onslow. (2009). Peer responses to stuttered utterances. *American Journal of Speech Language Pathology,* in press.

Lewis, Trent W., & David M. W. Powers. (2005). Distinctive feature fusion for improved audio-visual phoneme recognition. In Abdesselam Bouzerdoum & Azeddine Beghdadi (Eds.), *Proceedings of the 8th IEEE International Symposium on Signal Processing and Its Applications 2005* (pp. 62-65). Sydney, Australia: IEEE Press.

Lewis, Trent W., & David M. W. Powers. (2008). Distinctive feature fusion for recognition of Australian English consonants. *In Proceedings Interspeech 2008* (pp. 2671-2674). Brisbane.

Li, Yan, & David M. W. Powers. (2002). Speech separation based on higher order statistics using recurrent neural networks. In Ajith Abraham & Mario Köppen (Eds.), *Hybrid Information Systems, Proceedings of the 1st International Workshop on Hybrid Intelligent Systems* (pp. 45-56). Heidelberg: Physica-Verlag.

Loakes, Debbie, & Kirsty McDougall. (In press). Individual variation in the frication of voiceless plosives in Australian English: A study of twins' speech. *Australian Journal of Linguistics***.** Accepted January 2009.

McIntyre, Gordon, & Roland Goecke. (2007). Towards affective sensing. *Proceedings of the 12th International Conference on Human-Computer Interaction HCII2007*, *3*, 411-420.

McLeod, Sharynne, & David H. McKinnon. (2007). Prevalence of communication disorders compared with other learning needs in 14500 primary and secondary school students. *International Journal of Language and Communication Disorders, 42* (Supp. 1), 37-59.

Menzies, Ross G., Sue O'Brian, Mark Onslow, Ann Packman, Tamsen St Clare, & Susan Block. (2008). An experimental clinical trial of a cognitive behavior therapy package for chronic stuttering. *Journal of Speech and Hearing Research, 51,* 1451-1464.

Millar, J. Bruce, Phillip Dermody, Jonathan Harrington, & Julie Vonwiller. (1990). A national database of spoken language: Concept, design, and implementation. *Proceedings of the International Conference on Spoken*

Language Processing (pp. 1281-1284). Kobe, Japan: International Speech Communication Association, (ISCA) Archive, http://www.isca-speech.org/archive/icslp_1990

Mok, Mansze, David B. Grayden, Richard C. Dowell, & David Lawrence. (2006). Speech perception for adults who use hearing aids in conjunction with cochlear implants in opposite ears. *Journal of Speech, Language, and Hearing Research*, *49,* 338-351.

Naseem, Imran, Roberto Togneri, & Mohammed Bennamoun. (2009). Sparse representation for video-based face recognition. In Massimo Tistarelli & Mark S. Nixon (Eds.), *Advances in Biometrics: Third International Conference, ICB 2009* (pp. 219-228)., Alghero, Italy.

International Conference on Biometrics. Springer: Heidelberg.

National Indigenous Languages Survey Report. (2005). Department of Communications, Information Technology, Canberra.

Pfitzner, Darius M., Richard E. Leibbrandt, & David M. W. Powers. (2008). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems: An International Journal,* DOI 10.1007/s10115-008-0150-6

Pfitzner, Darius M., Kenneth Treharne, & David M. W. Powers. (2008). User keyword preference: The Nwords and Rwords experiments. *International Journal of Internet Protocol Technology*, 9, 149-158. DOI 10.1504/IJIPT.2008.020947

Potamianos, Gerasimos, Chalapathy Neti, Juergen Luettin, & Iain Matthews. (2004). Audio-visual automatic speech recognition: An overview. Electronic document available at: http://www.research.ibm.com/AVSTG/MITBOOK04_REVIEW.pdf Accessed on 30 July, 2009.

Powers, David M. W., & Richard E. Leibbrandt. (2009). Rough diamonds in natural language learning. Invited keynote (10pp), *Proc. Conference on Rough Sets and Knowledge Technology, Springer Lecture Notes in Computer Science (to appear).*

Powers, David M. W., Richard E. Leibbrandt, Darius M. Pfitzner, Martin H. Luerssen, Trent W. Lewis, Arman Abrahamyan, et al. (2008). Language teaching in a mixed reality games environment. *The 1st International Conference on Pervasive Technologies Related to Assistive Environments (PETRA)* DOI 10.1145/1389586.1389668

Rose, Philip. (2002). *Forensic speaker identification.* London: Taylor & Francis.

Ruben, Robert J. (2000). Redefining the survival of the fittest: Communication disorders in the 21st century. *Laryngoscope, 110,* 241-245.

Saragih, Jason, & Roland Goecke. (2007). A nonlinear discriminative approach to AAM fitting. *Proceedings of the Eleventh IEEE International Conference on Computer Vision ICCV2007* (pp. 14-20). Rio de Janeiro, Brazil: IEEE.

Smits, Roel, Natasha Warner, James M. McQueen, & Anne Cutler. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America, 113,* 563-574.

Talarico, Maria, Geraldine Abdilla, Martha Aliferis, Irena Balazic, Irene Giaprakis, Toni Stefanakis, et al. (2007). Effect of age and cognition on childhood speech in noise perception abilities. *Audiology & Neurotology*, *12*, 3-19.

Tran, Dat, & Michael Wagner. (2002). A fuzzy approach to speaker verification. *International Journal of Pattern Recognition and Artificial Intelligence*, *16* (7), 913-925

Tran, Dat, Michael Wagner, Yee W. Lau, & Mitsuo Gen. (2004). Fuzzy methods for voice-based person authentication. *Institute of Electrical Engineers of Japan Transactions on Electronics, Information and Systems*, *124* (10), 1958-1963.

Vidhyasaharan, Sethu, Eliathamby Ambikairajah, & Julien Epps. (2009). Speaker dependency of spectral features and speech production cues for automatic emotion classification. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 4693-4696). Taipei, Taiwan.

Viethen, Jette, & Robert Dale. (2006). Algorithms for generating referring expressions: Do they do what people do? *Proceedings of the International Conference on Natural Language Generation, Sydney, Australia* (pp. 63-72). Stroudsburg, PA, USA: The Association for Computational Linguistics.

Yacoub, Sherif, Steven Simske, Xiaofan Lin, & John Burns. (2003). Recognition of emotions in interactive voice response systems. pp. 729-732, *Proceedings of Eurospeech 2003 - 8th European Conference on Speech Communication and Technology , Geneva, Switzerland*, (pp. 729-732). International Speech Communication Association. ISCA Archive, http://www.isca-speech.org/archive/eurospeech_2003

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-435-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Burnham, Denis, Eliathamby Ambikairajah, Joanne Arciuli, Mohammed Bennamoun, Catherine T. Best, Steven Bird, Andrew R. Butcher, Steve Cassidy, Girija Chetty, Felicity M. Cox, Anne Cutler, Robert Dale, Julien R. Epps, Janet M. Fletcher, Roland Goecke, David B. Grayden, John T. Hajek, John C. Ingram, Shunichi Ishihara, Nenagh Kemp, Yuko Kinoshita, Takaaki Kuratate, Trent W. Lewis, Debbie E. Loakes, Mark Onslow, David M. Powers, Philip Rose, Roberto Togneri, Dat Tran, and Michael Wagner. 2009. A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 96-107. Somerville, MA: Cascadilla Proceedings Project.

or:

Burnham, Denis, Eliathamby Ambikairajah, Joanne Arciuli, Mohammed Bennamoun, Catherine T. Best, Steven Bird, Andrew R. Butcher, Steve Cassidy, Girija Chetty, Felicity M. Cox, Anne Cutler, Robert Dale, Julien R. Epps, Janet M. Fletcher, Roland Goecke, David B. Grayden, John T. Hajek, John C. Ingram, Shunichi Ishihara, Nenagh Kemp, Yuko Kinoshita, Takaaki Kuratate, Trent W. Lewis, Debbie E. Loakes, Mark Onslow, David M. Powers, Philip Rose, Roberto Togneri, Dat Tran, and Michael Wagner. 2009. A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 96-107. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2292.

Author/s:
Burnham, D; Ambikairajah, E; Arciuli, J; Bennamoun, M; Best, CT; Bird, S; Butcher, AR; Cassidy, S; Chetty, G; Cox, FM; Cutler, A; Dale, R; Epps, JR; Fletcher, JM; Goecke, R; Grayden, DB; Hajek, JT; Ingram, JC; Ishihara, S; Kemp, N; Kinoshita, Y; Kuratate, T; Lewis, TW; Loakes, DE; Onslow, M; Powers, DM; Rose, P; Togneri, R; Tran, D; Wagner, M

Title:
A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus

Date:
2009

Citation:
Burnham, D., Ambikairajah, E., Arciuli, J., Bennamoun, M., Best, C. T., Bird, S., Butcher, A. R., Cassidy, S., Chetty, G., Cox, F. M., Cutler, A., Dale, R., Epps, J. R., Fletcher, J. M., Goecke, R., Grayden, D. B., Hajek, J. T., Ingram, J. C., Ishihara, S. ,... Wagner, M. (2009). A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus. Selected Proceedings of the 2008 HCSNet Wokshop on Designing the Australian National Corpus: Mustering Languages, pp.96-107. Cascadilla Press.

Persistent Link:
http://hdl.handle.net/11343/32218

File Description:
Published