A Detailed Description of the AVOZES Data Corpus

Roland Goecke[1] **and J Bruce Millar**[1,2]

[1]National ICT Australia[†] [2]Australian National University
Canberra, Australia
roland.goecke@nicta.com.au bruce.millar@anu.edu.au

Abstract

The AVOZES data corpus has recently been made publicly available for other interested researchers. It is the first publicly available audio-video speech data corpus for Australian English. It contains recordings from 20 speakers and the sequences provide both a systematic coverage of the phonemes and visemes of Australian English as well as some application-driven utterances. AVOZES is also the first audio-video speech data corpus with stereo-video recordings, which enable a more accurate measurement of geometric facial features.

1. Introduction

Despite the growing interest that the field of Audio-Video Speech Processing (AVSP) has received in recent years, comprehensive, systematically designed audio-video (AV) speech corpora are still rare. However, such corpora form an important resource for the development of language-specific speech processing technology, e.g. automatic speech recognition (ASR) systems, and for the testing and comparison of results published by various research groups around the world. Although a number of corpora have been produced over the last few years, e.g. DAVID (Chibelushi et al. 1996), M2VTS (Messer et al. 1998), XM2VTSDB (Messer et al. 1999), Tulips1 (Movellan 1995), IBM LVCSR (Neti et al. 2000), CUAVE (Patterson et al. 2002), only some corpora appear to have been designed based on a comprehensive phonemic and visemic analysis, while the designs of others seem to be application-driven. Such an approach is perfectly acceptable for a particular project, but it limits the reusability of a corpus.

Given that the production of a speech data corpus is resource consuming task, the *A*udio-*V*ideo *OZ*stralian *E*nglish *S*peech (*AVOZES*) data corpus was designed and recorded with two major goals in mind, so that the corpus could serve more than one purpose and be useful to other researchers too. Firstly, AVOZES was to follow a modular framework for the recording of well-structured, comprehensive, multiple-use AV speech corpora as proposed in (Goecke et al. 2000) and recently generalised in (Millar et al. 2004). Secondly, AVOZES was to be made publicly available, as a foundational AV speech data corpus for Australian English (AuE) did not exist (the foundational ANDOSL data corpus (Millar et al. 1994) does not include a video component). In addition, AVOZES is novel in that its video recordings were made using a stereo camera system. A stereo vision system has the advantage over monocular systems that 3D coordinates can be recovered accurately. Thus, 3D distances can be measured, not just distances in 2D image coordinates, which makes the measurements robust against rotations of the face.

This paper expands a previous paper (Goecke and Millar 2004) and provides a detailed overview of the contents of AVOZES. Section 2. describes the design of the corpus. Section 3. gives details about the recording setup. The recording equipment is described in Section 4. An overview of the recorded data is presented in Section 5. Finally, further details are described in Section 6.

2. The AVOZES Design

The design of the AVOZES data corpus follows a modular framework proposed in (Goecke et al. 2000) which is also in accordance with the design methodology proposed in (Millar et al. 2004). A modular approach, where each module contains certain sequences, allows for extensibility in terms of the various factors that need to be addressed in corpus design (Chibelushi et al. 1996; Öhman 1998; Goecke et al. 2000). These include data collection factors (What are the conditions in the recording environment?), speaker factors (language background, speaking style, personal characteristics), and speech material factors (letters of the alphabet, isolated words, continuously spoken phrases; existing or nonsense words). The framework enables the design of AV speech corpora in a systematic way. The modular structure gives it the flexibility required to be useful for various research themes and applications, while the minimum requirements enable consistency across corpora.

In summary, the framework suggests that any AV speech data corpus contains three mandatory modules as a minimum, which cover the recording setup without and with speakers, as well as the actual speech material sequences which should contain the phonemes and visemes of a language. Additional optional modules can be added to cover specific issues, e.g. different view angles, different levels of illumination or acoustic noise.

AVOZES has a total of six modules — one general module and five speaker-specific modules. The modules are:

1. sampling the recording setup without a speaker,

For each speaker,

2. sampling the recording setup with speaker,
3. calibration sequences,

[†]National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

Class	IPA	“You grab ... beer.”	“as in ...”	Class	IPA	“You grab ... beer.”	“as in ...”
Oral stops	p	arpar	par	Short vowels	ɪ	bib	ship
	b	arbar	bar		ʊ	boub	should
	t	artar	tar		ɛ	beb	head
	d	ardar	dark		ə	bab *	the *(not “thee”)*
	k	arkar	car		ɒ	bob	shop
	ɡ	argar	garb		ʌ	bub	cup
Fricatives	f	arfar	far		æ	bab	had
	v	arvar	van	Long vowels	iː	beeb	heed
	θ	arthar	thin		uː	boob	cool
	ð	arthar	than		ɜː	berb	herb
	s	arsar	sue		ɔː	borb	floor
	z	arzar	zoo		ɑː	barb	hard
	ʃ	arshar	sure		əː	bareb	bare
	ʒ	arzjar *	azure	Diph-thongs	eɪ	babe	babe
	h	arhar *	ham		ɔɪ	boyb	boy
Affricates	tʃ	archar	chore		aɪ	bibe	hide
	dʒ	arjar	judge		aʊ	bowb	how
Nasals	m	armar	mow		ɪə	beerb	here
	n	arnar	now		əʊ	bobe	pope
	ŋ	arngar	sing		ʊə	boo-eb *	tour
Liquids and glides	l	arlar	lull				
	r	ara	row				
	w	arwar	wow				
	j	aryar	you				

Table 1: Phoneme classes and prompts for the AVOZES data corpus. Phonemes marked with an asterisk (*) were omitted from the recordings. IPA refers to the *International Phonetic Association* and its alphabet (IPA 1999).

4. short words in carrier phrase covering phonemes and visemes,
5. application sequences - digits, and
6. application sequences - continuous speech.

These modules are now explained in more detail.

2.1. Sampling Recording Setup without Speaker

Module 1 contains five sequences in AVOZES. The first two are 30s sequences of the recording scene viewed by the two cameras, but without any speaker present, one for each of the two recording periods (see Section 5.1.). The sequence can be used to determine the background level of acoustic noise present in the recording studio, due to air-conditioning as well as computer and recording equipment. In addition, information about the visual background can be gained, if it is required for the segmentation of the speaker from the background in the video stream. The other three sequences show a metronome in front of the two cameras, which provides information about the synchronisation of the audio and video streams (Robert-Ribes and Millar 1996). The sequences show the metronome on a slow setting, on a medium-paced setting, and on a fast setting.

Since the sequences in this module are speaker-independent, only one recording was needed. However, if corpus recordings were made over prolonged time spans (months or years), or in intervals (for example, extending the corpus at a later stage), the sequences should be repeated once during each interval to record possible changes to the recording environment.

2.2. Sampling Recording Setup with Speaker

Module 2 contains one sequence for each speaker showing horizontal head rotations. Such sequences can be useful for building face models, which are potentially not only of interest for AVSP but also for authentication purposes. The speaker is first shown in face frontal position for 5s, then the speakers turned their head 45° to the left, kept it there for 5s, then turned it 45° to the right of the frontal position and held that position for 5s again. These sequences are typically about 20s long.

2.3. Calibration Sequences

Module 3 comprises two sequences per speaker for the purpose of ‘speaker calibration’, in terms of their visible speech articulation or visual expressiveness. For (purely visual) lipreading as well as AV ASR, the amount of visible speech articulation determines how much (additional) information can possibly be gained from the video stream. A person with expressive visible speech articulation offers more information than one who does not move the visible speech articulators much (for example, a person who mumbles). Extracting lip parameters, such as mouth width or mouth height, over time enables an analysis of the visual expressiveness of a speaker, for example by analysing the maximum values reached in each cycle of lip movements. Speakers with values in the margin of the overall distribution can be treated differently, if desired.

The two calibration sequences “ba ba ba ...” (/bɑː bɑː bɑː .../) and “e o e o e o ...” (/iː ɔː iː ɔː iː ɔː .../) were each

Viseme Description	IPA
Bilabials	p b m
Labio-dentals	f v
Inter-dentals	θ ð
Labio-velar glides	w r
Palatals	ʃ tʃ ʒ dʒ
Alveolar non-fricatives and stops and velar stops	l n j h g k
Alveolar fricatives and stops	z s d t
Front non-open vowels and front close-onset diphthongs	iː ɪ ɛ ɪə
Open vowels and open-onset diphthongs	æ ɑː ɜː ʌ ə ɔː aɪ eɪ
Back/central non-open vowels and diphthongs	uː ʊ əː ɔɪ ʊə
Back/central open vowels and diphthongs	ɒ əʊ aʊ

Table 2: Viseme classes in Australian English.

repeated continuously by each speaker for about 10s. Despite the artificial nature of these prompts, the first sequence can give insight into the amount of vertical lip movement, i.e. opening and closing, while the second sequence emphasises horizontal lip movement, i.e. rounding and stretching.

2.4. Short Words in a Carrier Phrase Covering Phonemes and Visemes

The sequences in module 4 form the core part of AVOZES. There are 44 phonemes (24 consonantal and 20 vocalic phonemes) and 11 visemes (7 consonantal and 4 vocalic visemes) in AuE (Woodward and Barber 1960; Plant and Macrae 1977; Plant 1980). Following the ANDOSL design (Millar et al. 1994), the phonemes can be categorised into 8 classes (Table 1). Similarly, there are 11 viseme classes (Table 2)[1]. In (Plant 1980), it was also noted that the diphthong /aʊ/ appeared to be visually distinctive in a CVC-context[2] (with C=/b/), while this was not the case in the original study with a CV-context (with C=/b/). It might, therefore, be considered as an additional viseme.

The phonemes and visemes were put in central position in CVC- or VCV-contexts[3] to be free of any phonological or lexical restrictions. However, wherever possible, existing English words (that follow these context restrictions) were favoured over nonsense words in order to simplify the familiarisation process for the speakers. The vowel context for VCV-words was the wide open /ɑː/ (“ar-ar”). The voiced bilabial /b/ was used as the consonant context (“b-b”) for CVC-words. The opening and closing of a bilabial viseme clearly marks the beginning and end of the vocalic nucleus, thus facilitating the visual analysis. Using /b/ instead of /p/ lengthens each word, giving more data to analyse. A disadvantage of the /bVb/ context is that a bilabial context causes strong coarticulation effects in the formants. However, these are quite predictable for /b/ and we believe that the advantages of a bilabial context for visual segmentation outweigh the disadvantages from coarticulation.

To overcome the typical prosodic patterns associated with reading words from a list, each CVC- and VCV-word was enclosed by the carrier phrase “*You grab /WORD/ beer.*” Having a bilabial opening and closing before and after the word under investigation again helps with the visual segmentation process, in particular for the VCV-words. Table 1 shows the list of prompts and pronunciation hints, which were presented to the speakers during familiarisation and recording. Each phrase to be read aloud by the speakers was shown at the top of a prompt message on the screen, followed by an example of how to pronounce the phoneme in that prompt (see Figure 1 right).

Two phonemes from the list in Table 1 were omitted (marked with an asterisk (*)) because they have a low occurrence in AuE. These phonemes were /ʒ/ (as in “azure”) and /ʊə/ (as in “tour”). It was, therefore, considered to be likely that speakers would not pronounce the prompts correctly. These two phonemes were also rather difficult to achieve in the selected CVC- and VCV-contexts. Furthermore, the neutral vowel /ə/ and the neutral consonant /h/ were not recorded, because it was assumed at the time of recording that they add little to our understanding of AV relationships due to their neutrality.[4] During the recordings it also became evident, that some speakers had difficulties in producing distinguishable sounds for the voiceless and voiced inter-dental fricatives /θ/ and /ð/, as well as producing the velar closure nasal /ŋ/. The analysis of these sequences must therefore be treated with care. The AVOZES documentation contains a list of sequences in question (Goecke 2004).

2.5. Application Sequences - Digits

Digit recognition is a common task in ASR research and similar sequences can be found in a number of AV speech corpora (e.g. DAVID (Chibelushi et al. 1996), Tulips1 (Movellan 1995)). AVOZES includes one sequence per digit for each speaker, spoken in order from 0 to 9. Each digit is enclosed by the carrier phrase “*You grab /DIGIT/ beer.*” to ensure lip closure before and after the digit for ease of segmentation of the video stream.

2.6. Application Sequences - Continuous Speech

This second module with application-driven sequences contains examples of continuous speech from each speaker. The three sequences are:

1. “*Joe took father’s green shoe bench out.*”[5]
   /dʒəʊ tʊk fɑːðɛz griːn ʃuː bɛntʃ aʊt/

[1] The phonemes /z/, /ʒ/, /h/, and /ŋ/ were not included in the investigation by Plant and Macrae, but are here classified into corresponding viseme classes in Table 2.

[2] CVC - consonant-vowel-consonant

[3] VCV - vowel-consonant-vowel

[4] In hindsight, it might have been better to also record these four phonemes at the time for completeness, even if speakers had difficulties producing the correct pronunciation. However, these sequences could be recorded and added to AVOZES in future, due to the modular design of the data corpus.

[5] This sentence appeared also in the corpora M2VTS and XM2VTSDB (Messer et al. 1999).

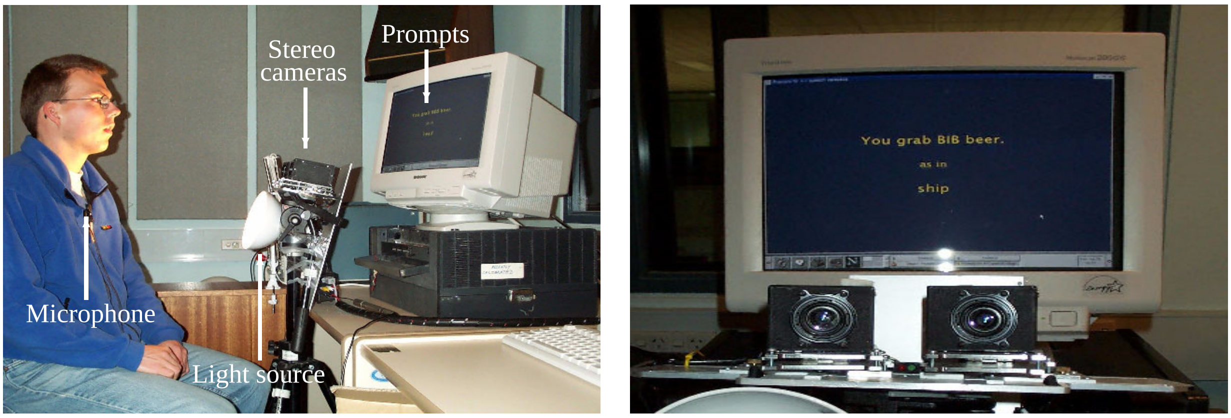

Figure 1: Left: Recording setup. Right: Speaker's view of the recording setup and the prompts on screen.

2. "*Yesterday morning on my tour, I heard wolves here.*"
   /jɛstədeɪ mɔːnɪŋ ɒn maɪ tʊə aɪ hɜːd wʊlvz hɪə/

3. "*Thin hair of azure colour is pointless.*"
   /θɪn həː ɒv eɪʒə kʌlə ɪs pɔɪntləs/

Together with the first sentence, the second and third sentences were designed in such a way that they contain almost all phonemes and visemes of AuE (/æ/ is the only phoneme missing). One of the ultimate goals in automatic speech recognition is the task of continuous speech recognition in all conditions. The sequences in this module offer an initial way of applying and testing any results from an analysis of the sequences in module 4 to such a task.

3. Recording Setup

3.1. Recording Studio Layout

The recording of the AVOZES data corpus took place in the audio laboratory of the Computer Sciences Laboratory, RSISE, at the Australian National University. The same equipment was used on both occasions. The laboratory is a soundproof room in the interior of the building, well-shielded from noise sources outside the room but with a small amount of background acoustic noise from the room's air-conditioning and the recording equipment. The latter was housed in an acoustically insulated box.

Figure 1 left shows the recording setup. The speakers sat on an office swivel chair in front of the stereo cameras, which were positioned with the help of a camera tripod. A light source was placed directly below the camera rig to illuminate the speaker's face, while blinding was reduced to a minimum. Other light source arrangements were considered, but were discarded in favour of the simplicity of a single light source. In addition, the scene was illuminated from three ceiling mounted incandessant lights.

An office swivel chair was used for two reasons. Firstly, the height of the seat could be adjusted easily for shorter or taller people, while leaving other equipment unchanged. Secondly, the sequences in module 2 (Section 2.2.) require the speaker to turn the head $45°$ to the left and to the right of the cameras. By sitting on a swivel chair, the speakers could simply turn their whole bodies towards markers on the wall. Keeping the vertical axis of the chair aligned with the torso at a marked position ensured that the face was kept in the cameras' viewfields. The distance from the face to the cameras was 600 ± 50mm, which corresponded to the distance ("depth") range that the cameras were calibrated for. Speakers were allowed to move their head freely, but were asked to keep it roughly in the same position to ensure that it was within the cameras' viewfield.

3.2. Prompts

The speaking prompts appeared on a computer screen, which was about 200mm (in horizontal direction) behind the cameras (Figure 1). Prompts were advanced per mouse click by the recording assistant, when a prompt was pronounced correctly. Otherwise, the speaker was asked to repeat the phrase. The screen's background colour was swapped between a dark green and a dark blue whenever the next prompt appeared, giving the speaker an additional visual signal that a new prompt had appeared on the screen.

4. Recording Equipment

A clip-on microphone was attached to the speaker's clothes on the chest about 20cm below the mouth. The microphone was an omnidirectional Sennheiser MKE 10-3 microphone with a frequency response of 50Hz–20kHz. The microphone was directly connected to the Digital Video (DV) recorder, where the microphone's output was recorded on DV tape. The recorder was a JVC HR-DVS1U miniDV/S-VHS video recorder. DV recordings offer a good compromise between easy and fast access to the data and high-quality, yet inexpensive storage.

The two cameras were standard, colour analogue NTSC cameras mounted side by side on a rig. The cameras were placed on the rig with a slight vergence ($\approx 5°$) towards the centre. The output of the stereo cameras was multiplexed into one video signal using field multiplexing (Matsumoto et al. 1997), before being sent to the DV recorder. Images from both cameras can thus be stored in a single video frame and stereo image processing can be performed without any special hardware. Due to this setup, the video stream lags the audio stream by one video frame, which must be taken care of in any analysis of the data.

A weakness of field multiplexing is that only half the vertical resolution of the original video frame from each camera is available, as two video streams are compressed

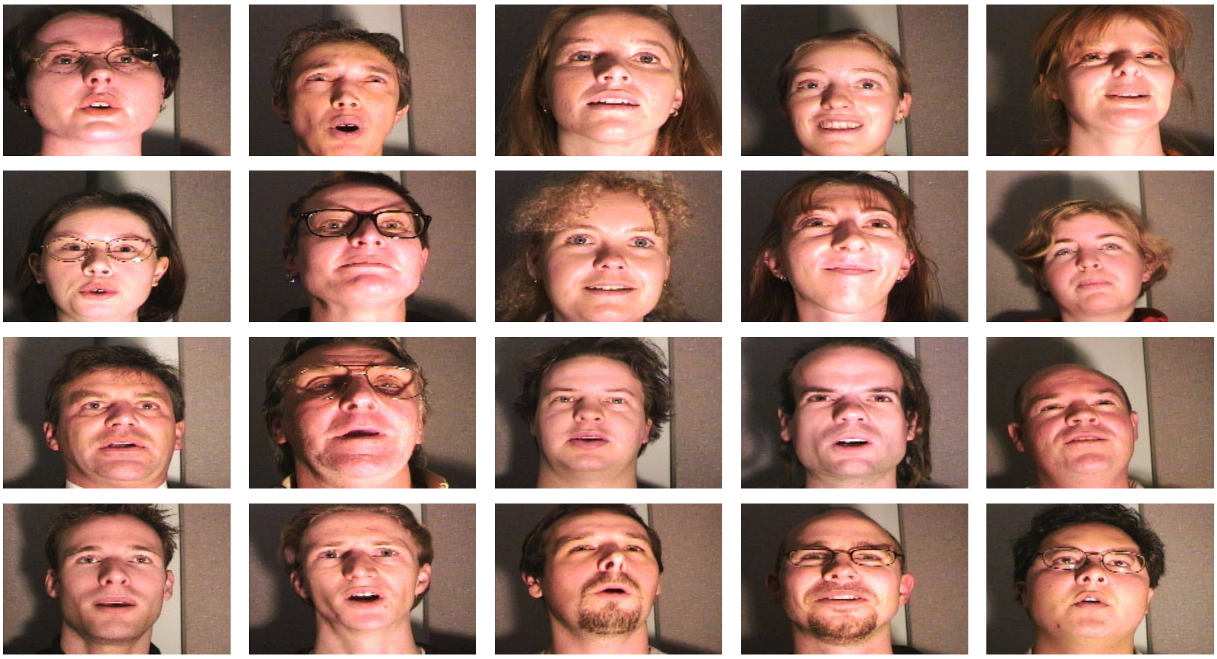

Figure 2: Face shots of the 20 native speakers of AuE in AVOZES.

into a single frame. Future studies should consider rotating the stereo cameras by 90°, so that the halved resolution is in the horizontal direction rather than the vertical direction, which is potentially the more informative axis in visible speech articulation. One other weakness is the delay between the two images in each video frame, which is inherent in any interlaced video/TV standard. That is, first all the lines of one field are processed, then all the lines of the other field. The field frequency is 60Hz in the NTSC standard and hence there is a 16.6ms delay between fields and in this case camera outputs. In the authors' experience, this delay has not posed a problem in the analysis of the data.

The cameras were calibrated, so that 3D coordinates could be recovered, which in turn enables 3D measurements of distances on a speaker's face. The calibration process and its results have been described in the AVOZES documentation (Goecke 2004).

5. Recorded Data

5.1. Sequences

AVOZES currently contains recordings made from 20 native speakers of AuE (10 female + 10 male speakers). Six speakers wear glasses, three wear lip make-up, two have beards (Figure 2). At the time of the recordings, these speakers were between 23 and 56 years old. The speakers were tentatively classified into the three speech varieties of AuE (broad, general, cultivated) by the recording assistant, which created groups of 6 speakers for broad AuE, 12 speakers for general AuE, and only 2 speakers for cultivated AuE. While this distribution approximately reflects the composition of the Australian population in terms of accent varieties, it should be noted that individual groups are not gender balanced, and that their size is small for statistical analyses on an individual group basis. It is also worthwhile to remember that the accent varieties are not discrete entities, but rather span a continuum of accent variation, so that some classifications represent a best estimate.

Recordings were made at two occasions. The first set of 10 native speakers of AuE was recorded over the period of one week in August 2000. The second set of another 10 native speakers of AuE was recorded over a period of two days in August 2001, using exactly the same equipment, setup, and location as in the first set.

Each speaker spent about 30min in the recording studio. They were first familiarised with the speech material and informed about the recording procedure. Actual recordings took about 5min per speaker. A total of 56 sequences were recorded per speaker (1 face sequence and 55 speech material sequences) with no repetitions. A recording assistant was present, so that speakers did not have to handle any of the equipment themselves and could concentrate on the speaking task. The AVOZES data corpus currently contains only frontal face ($\pm 10°$) AV speech recordings, with no separate or simultaneous recordings from a different angle. Recordings were made for a clean audio condition.

5.2. Speaker Data

Beside the actual recordings, each speaker provided personal data, to enable such data to be used in the interpretation of the analysis of the recorded material. Personal data collected contains:

- date of birth, and gender,
- level of education and current occupation,
- height and weight,
- native language of speaker, speaker's mother, and speaker's father,
- place of origin and occupation of both parents,

- extended periods outside Australia (at least 3 months) — time and place,
- singing, training in singing,
- smoking, medical conditions (e.g. asthma).

In addition, the distance from the speaker's mouth to the microphone was also measured. The individual information (names omitted, date of birth transformed into age) about the native speakers of AuE is presented in the AVOZES documentation (Goecke 2004).

6. Summary

In total, AVOZES offers about 20GB of AV data and is distributed on DVDs. All sequences are available as AVI files containing both audio and video information as well as WAV files containing only the audio component. Video information is encoded using the NTSC YUV 4:1:1 format, 720×480 pixels, 29.97Hz frame rate. The AVOZES AVI files use the Adaptec DVSoft codec, which most media players support. Audio information is encoded as 48kHz, 16-bit stereo (the two stereo channels contain the same information as only a mono microphone was used). The length of the individual sequences has been chosen to be a multiple of a full second (i.e. of 30 video frames).

The authors hope that AVOZES will be useful for other researchers and that it marks the starting point for interest in a larger AV speech data corpus for AuE, perhaps supported through the *Enabling Human Communication* ARC research network. Readers interested in acquiring a licence should contact the authors of this paper. Further details about AVOZES and its licence terms can be found at <http://rsise.anu.edu.au/∼roland/>.

References

Chibelushi, C., S. Gandon, J. Mason, F. Deravi, and D. Johnston (1996). Design Issues for a Digital Integrated Audio-Visual Database. In *IEE Coll. Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, London, UK, Digest Reference Number 1996/213, pp. 7/1–7/7.

Goecke, R. (2004). The Audio-Video Australian English Speech Data Corpus AVOZES. Documentation, Australian National University.

Goecke, R. and B. Millar (2004). The Audio-Video Australian English Speech Data Corpus AVOZES. In *Proc. 8th Int. Conf. Spoken Language Processing ICSLP2004*, Volume III, Jeju, Korea, pp. 2525–2528.

Goecke, R., Q. Tran, B. Millar, A. Zelinsky, and J. Robert-Ribes (2000). Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing. In *Proc. 8th Australian Int. Conf. Speech Science and Technology SST2000*, Canberra, Australia, pp. 92–97.

IPA (1999). *Handbook of the International Phonetic Association*. Cambridge, UK: Cambridge Univ. Press.

Matsumoto, Y., T. Shibata, K. Sakai, M. Inaba, and H. Inoue (1997). Real-Time Color Stereo Vision System for a Mobile Robot based on Field Multiplexing. In *Proc. IEEE Int. Conf. Robotics and Automation ICRA'97*, Albuquerque (NM), USA, pp. 1934–1939.

Messer, K., J. Matas, and J. Kittler (1998). Acquisition of a large database for biometric identity verification. In *Proc. BIOSIGNAL 98*, Brno, Czech Republic, pp. 70–72.

Messer, K., J. Matas, J. Kittler, J. Luettin, and G. Maitre (1999). XM2VTSDB: The Extended M2VTS Database. In *Proc. 2nd Int. Conf. Audio and Video-based Biometric Person Authentication AVBPA'99*, Washington (DC), USA, pp. 72–77.

Millar, B., J. Vonwiller, J. Harrington, and P. Dermody (1994). The Australian National Database Of Spoken Language. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing ICASSP'94*, Volume 1, Adelaide, Australia, pp. 97–100.

Millar, B., M. Wagner, and R. Goecke (2004). Aspects of Speaking-Face Data Corpus Design Methodology. In *Proc. 8th Int. Conf. Spoken Language Processing ICSLP2004*, Volume II, Jeju, Korea, pp. 1157–1160.

Movellan, J. (1995). Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen (Eds.), *Advances in Neural Information Processing Systems*, Volume 7, Cambridge (MA), USA, pp. 851–858. MIT Press.

Neti, C., G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou (2000). Audio-Visual Speech Recognition. Workshop report, CSLP / Johns Hopkins University, Baltimore, USA.

Öhman, T. (1998). An audio-visual speech database and automatic measurements of visual speech. Quarterly Status and Progress Report TMH-QPSR 1-2/1998, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Patterson, E., S. Gurbuz, Z. Tufekci, and J. Gowdy (2002). CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing ICASSP2002*, Volume 2, Orlando (FL), USA, pp. 2017–2020.

Plant, G. (1980). Visual identification of Australian vowels and diphthongs. *Australian Journal of Audiology 2*(2), 83–91.

Plant, G. and J. Macrae (1977, July). Visual Perception of Australian Consonants, Vowels and Diphthongs. *Australian Teacher of the Deaf 18*, 46–50.

Robert-Ribes, J. and B. Millar (1996). A simple system for measuring audiovisual speech. In *Proc. 6th Australian Int. Conf. Speech Science and Technology SST-96*, Adelaide, Australia, pp. 13–18.

Woodward, M. and C. Barber (1960, September). Phoneme Perception in Lipreading. *Journal of Speech Hearing Research 3*(3), 212–222.