The Audio-Video Australian-English Speech Data Corpus AVOZES

Documentation Version 1.2

Roland Göcke

Email: roland.goecke@ieee.org

Mail: c/o National ICT Australia
Canberra Laboratory
Locked Bag 8001
Canberra ACT 2601
Australia

Canberra, 14 November 2004

Copyright Notice

Contents

Chapter 1

Introduction

For testing and comparing results published by various research groups in the field of AVSP, a common basis in the form of a comprehensive, systematically designed AV speech data corpus would be of great value. Such a publicly available 'benchmark' AV speech data corpus still does not exist, despite a number of corpora having been produced over the last few years. Many corpora appear to have been designed with a specific application in mind, rather than being based on a general phonemic and visemic analysis.

The *A*udio-*V*ideo *OZ*stralian *E*nglish *S*peech (*AVOZES*) data corpus was designed and recorded with two major goals in mind. Firstly, a new framework for the design of comprehensive, well-structured, multiple-use AV speech data corpora was proposed and followed in the production of the AVOZES data corpus. Secondly, the first publicly available, comprehensive AV speech data corpus for Australian English (AuE) was produced. In addition, it is the first AV speech data corpus to use a stereo vision 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 document is meant to serve as a reference documentation for using the AVOZES data corpus. It provides background information on AV speech data corpora in general and detailed information about what data AVOZES can offer. If you have received this document as part of a licence agreement, please feel free to contact the author with suggestions for improving this document and possibly AVOZES. If you have downloaded this document because you are interested in finding out more detail about the data available in AVOZES, I hope you find this document useful. If you would like to acquire a licence for AVOZES, please contact the author (contact details can be found on the title page).

Chapter 2

Some Other AV Speech Corpora

This section gives an overview of some other AV speech data corpora commonly used in AVSP for comparison with the AVOZES data corpus. There is also some overlap with corpora which are mainly used for AV biometric person authentication. Only some major corpora for the English language are discussed here.

A good overview of existing AV speech corpora is given by Chibelushi *et al.* [2]. Their study led to the creation of the well-designed DAVID corpus [3] which consists of four different subcorpora, each addressing a particular research issue. The first subcorpus addresses the issue of facial image segmentation under different conditions, including variable illumination, variable backgrounds, and facial distractors such as glasses and hats. This subcorpus has 6 subjects. The second subcorpus is designed for research in the area of automatic speech and person recognition and contains recordings of 31 clients and 92 impostors. A subset of 9 subjects has highlighted lips (blue make-up) to facilitate the lip extraction process. Both the first and second subcorpus use the set of digits from 0 to 9 as speech material. Subcorpus 3 is intended for speech-assisted video compression and the synthesis of talking heads. VCVCV[1] utterances of 5 subjects were recorded. The fourth subcorpus is concerned with automatic speech and person recognition with application in video-conferencing systems. Hence, it contains sentences from a business control set spoken by 31 clients and 92 impostors. All recordings show a frontal and profile view, achieved by a mirror construction and a single camera, together with the associated synchronous audio.

A well-established AV speech data corpus is the M2VTS database and its

[1] VCVCV = Vowel-Consonant-Vowel-Consonant-Vowel

successor XM2VTSDB [18, 19]. Whereas the M2VTS database contains 37 speakers, the XM2VTSDB database comprises recordings of 295 speakers. Four sessions were recorded to account for natural changes in appearance of the speakers. During each session, an AV speech recording was made as well as a head rotation sequence. The speech material recorded consists of three sequences, two of which contain the digits from 0 to 9 in different order. The third sequence is *"Joe took father's green shoe bench out."* which was designed to maximise visible articulatory movements. It contains all phoneme and viseme categories (but not all phonemes). The XM2VTSDB is currently the largest publicly available AV corpus in terms of numbers of speakers but suffers from the small number of different sequences for each speaker with respect to a complete phonemic and visemic analysis.

The Tulips1 data corpus recorded by Movellan [22, 23] contains the four digits 'one', 'two', 'three', and 'four' repeated twice by 9 male and 3 female subjects. This speech material was chosen with a phone number spelling task in mind. Only frontal views are recorded. As such, the corpus is rather small and application-driven.

Other AV speech databases have been recorded by various research groups but are not publicly available. One such proprietary data corpus is the IBM LVCSR[2] AV corpus [24], which contains continuously spoken utterances of the IBM ViaVoice training set from more than 290 American English speakers in different environments (office, car). The video stream is compressed using MPEG-2.

Recently, the CUAVE corpus was introduced by Patterson *et al.* [29]. It contains recordings from about 50 American English speakers, uttering connected and isolated digits. The sequences are stored as MPEG-2 files. The data is fully labelled at a millisecond level.

Although a comprehensive and systematically-designed audio data corpus exists for AuE (ANDOSL [20]), no AV speech corpora exist. As a result, AVOZES has been created as a new AV corpus for AuE. The next sections provide details about the data available in AVOZES.

[2]LVCSR = Large Vocabulary Continuous Speech Recognition

Chapter 3

The Design Framework

The design of the AVOZES data corpus followed a newly proposed, extensible framework, covering all visemes and almost all phonemes of AuE (Section 4). It is thereby the most comprehensive AV speech data corpus of AuE to date. This section provides some background information and consideration for the framework, followed by the framework itself.

3.1 What Do Researchers Look for in an AV Speech Data Corpus?

A survey by Chibelushi *et al.* [2] examined existing AV speech data corpora as well as which features researchers would like to see in such corpora. The latter was established by a questionnaire-style survey. Although only five questionnaires were received in response, the conclusions that were made by the authors match the observations made by the author of this document. The usual limitations of existing AV speech corpora are:

- a small number of speakers,
- a small number of phonemes and visemes covered, and
- isolated words such as digits or letters of the alphabet rather than embedded (carrier phrase) or continuous speech.

Most existing AV speech corpora were clearly designed for a particular research project and not as a publicly available corpus for the comparison of methods developed by various research groups around the world. The limited size of many corpora is clearly related to the time- and resource-consuming effort required in the creation of a corpus.

The features that researchers would like to see in a benchmark corpus were:

- a large number of speakers for statistical significance,
- a broad coverage of phonemes and visemes,
- different levels of acoustic noise starting with 'clean speech' (no noise),
- whole face images in colour,
- short words and continuous speech with transcription, and
- extensibility.

Chibelushi *et al.* [3] presented a number of ideas on how to design an AV speech data corpus that covers a variety of experimental themes. Similar considerations on design issues were made by Öhman [28] for a Swedish language AV speech data corpus. A general discussion of aspects in AV speech data corpus design can be found in Millar *et al.* [21]. Generally, various factors play a role in the design of such corpora, and these are briefly discussed in the next section.

3.2 Factors in AV Speech Corpus Design

3.2.1 Data Collection Factors

These factors relate to the corpus recording process. One can argue that recordings made in laboratories do not mirror exactly the conditions in the real world. However, in terms of facilitating the interpretation of experimental results, it is an advantage to be able to control the experimental conditions. These conditions include the recording equipment, the possible use of markers, the layout of the recording room (e.g. background), the sitting arrangement, the illumination arrangement, and the level of acoustic noise. Going through all possible combinations of these conditions in a systematic way would result in an exponential growth of the corpus and quickly become impractical. It is suggested here to leave all conditions but one constant at a time, and to study the effects of changing that condition, rather than mixing the effects of various changing conditions in one recording.

3.2.2 Speaker Factors

Speaker-related factors can be categorised into language background, speaking style, and personal characteristics. The first category includes issues like dialects (or accents) and first versus other languages. Usually, one would study native speakers first to characterise a particular language, but the identification of differences between native and foreign speakers is also an interesting research topic. Within a language, different dialects exist even among native speakers, and these must be considered. The second category, speaking style, determines the social and / or emotional conditioning of speaking, for example a conversational style or an excited style. It should be noted that first-time participants in a data corpus collection often feel nervous about the task ahead or are overly motivated to do the task particularly well, so that their speaking style changes from the way they would normally speak. Therefore, the familiarisation of speakers with the environment and the speech material is important. The third category deals with the 'natural' characteristics of a person, such as gender, body physique, characteristics of the vocal tract, or the amount of visible movement of the articulators.

Generally, a balanced population in a corpus is desirable. Finding a sufficient number of speakers to cover normal variation in the above categories might not always be possible, but it is suggested here to at least achieve a gender balance in groups with the same language background (native speakers, foreign speakers). As mentioned for the data collection factors, a systematic way of going through all possible combinations of these factors leads to an exponential growth of the corpus. It is, therefore, necessary to define the range of speaker-related factors, which are addressed in a corpus, in advance and to select speakers accordingly.

3.2.3 Speech Material Factors

These factors relate to the material that speakers are requested to speak. Such material can be letters of the alphabet, isolated words, and continuously spoken phrases. Words can be real, existing words (e.g. digits, commands) or nonsense words, designed to investigate a particular phoneme transition (e.g. ABABA). Which material is included in a particular corpus depends on the application in mind. It is suggested here that general-purpose corpora contain some examples from all categories. It should be noted that reading lists of phones, diphones, words etc. often results in a speaking style different from what it would be, if the words were included in a phrase. The use of a carrier phrase, in which the speech material unit of interest is embedded, is therefore suggested, unless the target application in mind requires otherwise

(for example, a spelling task). This problem does not occur (or only to a much smaller extent) for continuously spoken phrases.

Furthermore, the coverage of phonemes and visemes is another factor. Whenever feasible, it is suggested to cover all phonemes and visemes of a language at least once in the chosen context (phones, diphones, words etc.) for completeness. That means that except for very large corpora, not all possible diphone or triphone transitions will be covered. By using the same context, the phonemes and visemes can at least be studied in a controlled environment. Moreover, if the resources do not allow the inclusion of each phoneme, a careful selection must take place, so as to choose at least one for each viseme category, as the number of visemes is smaller than the number of phonemes (see Section 4.4).

3.3 The Proposed Framework

The ideas presented in Section 3.2 are extended here and a new framework for the design of AV speech data corpora is proposed. This framework is in accordance with the design methodology proposed by Millar *et al.* [21]. A modular approach, where each module contains certain sequences, allows for extensibility in terms of the various factors discussed in the previous subsection. For example, a data corpus could start with a small number of speakers uttering selected phoneme sequences in a noiseless audio condition. Later, more sequences can be recorded to extend the phonemic coverage, add more speakers, or repeat sequences in different noise levels. Thus, a corpus can grow over time, thereby accommodating the amount of resources it takes to create and store it, while still providing usable data from the beginning. In this context, ensuring continuity in the facilities and equipment used, as well as in terms of the speakers appearing in the corpus, is important. If recordings are made at different points in time, the comparability of the recorded material with earlier recordings is an important issue that needs to be addressed.

As a minimum, any AV speech data corpus should contain the following three modules:

1. sampling recording setup without a speaker,

For each speaker,

2. sampling recording setup with speaker, and
3. recording of phonemes and visemes.

The module "sampling recording setup without a speaker" captures general aspects of the data collection process, such as visual background, scene illumination, and acoustic background. For every speaker, there are at least two modules. The module "sampling recording setup with speaker" shows the speaker in the scene. This can include sequences useful or necessary for the video processing, such as shots of the face from various angles. The module "recording of phonemes and visemes" contains the actual speech material sequences following the guidelines in the previous section.

Additional modules can be added easily. Some modules, for each speaker, that were considered prior to the creation of the AVOZES data corpus, were:

- speaker calibration,
- application sequences,
- different view angles,
- different levels of illumination, and
- different levels of acoustic noise.

The module "speaker calibration" could contain sequences which exhibit particular acoustic or visible speech patterns (for example, lip rounding). These sequences can be used to classify speakers into different classes in the analysis stage. Longer sequences of continuous speech or command sequences would make up the module "application sequences". The other three modules comprise changes in data collection factors. From a data analysis point of view, repeating the modules "recording setup with speaker" and "coverage of phonemes and visemes" for each different condition is desirable. However, it must be noted that this may not be practically feasible, both in terms of the amount of resources and the duration of session times required, if a lot of speech material has to be covered. Speakers get tired if recording sessions become too long, so either the amount of speech material must be reduced, that is, not to cover all phonemes and visemes, or recordings must be made in different sessions, which raises questions of the comparability of the recordings because a speaker's mood or health might have changed between sessions. The longer the time span between sessions, the more pertinent these questions become.

The proposed 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 help to achieve consistency across corpora.

Chapter 4

Design of the AVOZES Data Corpus

The proposed framework was followed in the design of the *A*udio-*V*isual *OZ*stralian *E*nglish *S*peech (*AVOZES*) data corpus [8, 9]. No other AV speech data corpus with stereo camera video has been published thus far.

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

1. sampling recording setup without a speaker,

For each speaker,

2. sampling recording setup with speaker and definition of face model,
3. calibration sequences,
4. short words in carrier phrase covering phonemes and visemes,
5. application sequences - digits, and
6. application sequences - continuous speech.

These modules are described in more detail in the following sections.

4.1 Module 1 — Sampling Recording Setup without Speaker

This module contains five sequences in the AVOZES data corpus. The first two sequences (`backgroundNoiseWithoutSpeaker30s_1.avi` and `backgroundNoiseWithoutSpeaker30s_2.avi`) are 30 second sequences of the recording scene viewed by the two cameras, but without any speaker present,

one for each recording period (see Sections 5.1 and 7.1). The sequences 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 in this module show a metronome in front of the two cameras, which provides information about the synchronisation of the audio and video streams. Sequence `slowMetronome20s.avi` shows the metronome on a slow setting, sequence `moderateMetronome30s.avi` shows it on a medium-paced setting, and sequence `fastMetronome30s.avi` shows it on a fast setting. **It must be noted, that due to the stereo camera and recording setup, there is a delay of one NTSC video frame between audio and video streams (video lagging the audio, see Section 5 for details)**.

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.

4.2 Module 2 — Sampling Recording Setup with Speaker

Module 2 contains one sequence for each speaker showing sideway head movements. 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 5 seconds, then the speakers turned their head 45° to the left (as viewed by the speaker), kept it there for 5 seconds, then turned it 45° to the right of the frontal position and held that position for 5 seconds again. These sequences are typically about 20 seconds long.

4.3 Module 3 — Calibration Sequences

This module 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 automatic speech recognition, the amount of visible speech articulation determines how much (additional) information can possibly be gained from the video stream. Expressive

visible speech articulation offers more information than a person 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 excluded from the analysis or 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 .../) recorded in the AVOZES data corpus were each repeated continuously by each speaker for about 10 seconds. 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.

4.4 Module 4 — Short Words in a Carrier Phrase Covering Phonemes and Visemes

The sequences in this module form the core part of the AVOZES data corpus. They were initially intended for statistical analyses of relationships between audio and video speech parameters, but can of course be used for other goals, too. There are 44 phonemes (24 consonantal and 20 vocalic phonemes) and 11 visemes (7 consonantal and 4 vocalic visemes) in AuE, according to Woodward and Barber [37], Plant and Macrae [32], and Plant [31]. Following the ANDOSL design [20], the phonemes can be categorised into 8 classes (Tables 4.1 and 4.2). Similarly for the visemes, following [32] and [31], there are 11 viseme classes (Table 4.3)[1]. Plant and Macrae [32] do not label their vowel and diphthong visemes, but they are broadly:

1. front non-open vowels and front close-onset diphthongs,

2. open vowels and open-onset diphthongs,

3. back/central non-open vowels and diphthongs containing these vocalic positions, and

4. back/central open vowels and diphthongs containing these vocalic positions.

[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 4.3.

Class	Description	IPA Symbol	Example "as in ..."
Oral stops	Bilabial voiceless	p	poor
	Bilabial voiced	b	bore
	Alveolar voiceless	t	tore
	Alveolar voiced	d	door
	Velar voiceless	k	core
	Velar voiced	g	gore
Fricatives	Labio-dental voiceless	f	fan
	Labio-dental voiced	v	van
	Inter-dental voiceless	θ	thin
	Inter-dental voiced	ð	than
	Alveolar voiceless	s	sue
	Alveolar voiced	z	zoo
	Palatal voiceless	ʃ	sure
	Palatal voiced	ʒ	azure
	Glottal voiceless	h	ham
Affricates	Alveolar voiceless	tʃ	chore
	Alveolar voiced	dʒ	judge
Nasals	Bilabial closure	m	mow
	Alveolar closure	n	now
	Velar closure	ŋ	sing
Liquids and glides	Lateral	l	lull
	Rhotic	r	row
	Bilabial	w	wow
	Palatal	j	you

Table 4.1: Consonant phoneme classes in the ANDOSL and AVOZES data corpora. IPA refers to the *International Phonetic Association* and its alphabet. The latest version was published in 1993 and updated in 1996 [11].

In [31], these visemes were described in terms of their mouth shape as (1) small aperture and spread lips, (2) large aperture and neutral lips, (3) small aperture and rounded lips, and (4) large aperture and rounded lips. 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.

[2]CVC - consonant-vowel-consonant

Class	IPA Symbol	Example "as in ..."
Short vowels	ɪ	hid
	ʊ	hood
	ɛ	head
	ə	the *(not "thee")*
	ɒ	hod
	ʌ	bud
	æ	had
Long vowels	iː	heed
	uː	who'd
	əː	there
	ɜː	heard
	ɔː	hawed
	ɑː	hard
Diphthongs	eɪ	hay
	əʊ	hoed
	ɔɪ	hoy
	aɪ	hide
	aʊ	how
	ɪə	here
	ʊə	tour

Table 4.2: Vowel phoneme classes in the ANDOSL and AVOZES data corpora.

The phonemes and visemes in the AVOZES data corpus 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 of the speakers with the speech material. 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, and thus facilitates 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 pre-

[3]VCV - vowel-consonant-vowel

Viseme Description	IPA Symbols
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 plosives and velar plosives	l n j h g k
Alveolar fricatives and plosives	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 4.3: Viseme classes in Australian English.

dictable for /b/ and it is believed that the advantages of a bilabial context for visual segmentation outweigh the disadvantages from coarticulation.

To overcome the typical articulation 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. Tables 4.4 and 4.5 show the lists of prompts and pronunciation hints, which were presented to the speakers during familiarisation and recording. Each phrase to be read out aloud by the speakers was shown at the top of the prompt message on the screen, and was followed by an example of how to pronounce the phoneme under investigation in that prompt. For an example of such a prompt message, see Figure 5.2 in Section 5.

Two phonemes from the lists in Tables 4.1 and 4.2 were omitted (see also prompt lists in Tables 4.4 and 4.5; omitted phonemes are 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

Class	IPA Symbol	"You grab ... beer."	Pronunciation "as in ..."
Short vowels	ɪ	bib	ship
	ʊ	boub	should
	ɛ	beb	head
	ə	bab *	the *(not "thee")*
	ɒ	bob	shop
	ʌ	bub	cup
	æ	bab	had
Long vowels	iː	beeb	heed
	uː	boob	cool
	ɜː	berb	herb
	ɔː	borb	floor
	ɑː	barb	hard
	əː	bareb	bare
Diph-thongs	eɪ	babe	babe
	ɔɪ	boyb	boy
	aɪ	bibe	hide
	aʊ	bowb	how
	ɪə	beerb	here
	əʊ	bobe	pope
	ʊə	boo-eb *	tour

Table 4.4: Prompts for vowels and diphthongs in the AVOZES data corpus. Phonemes marked with an asterisk (*) were omitted from the recordings.

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 that they add little to the statistical analysis of relationships between audio and video speech parameters due to their neutrality.[4] During the recordings it also became evident, that some speakers had difficulties in producing distinguished 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.

[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 can and may be added to the AVOZES data corpus in future, due to the modular design of the data corpus.

Class	IPA Symbols	"You grab ... beer."	Pronunciation "as in ..."
Oral stops	p	arpar	par
	b	arbar	bar
	t	artar	tar
	d	ardar	dark
	k	arkar	car
	g	argar	garb
Frica-tives	f	arfar	far
	v	arvar	van
	θ	arthar	thin
	ð	arthar	than
	s	arsar	sue
	z	arzar	zoo
	ʃ	arshar	sharp
	ʒ	arzjar *	azure
	h	arhar *	hard
Affri-cates	tʃ	archar	chart
	dʒ	arjar	jar
Nasals	m	armar	arm
	n	arnar	barn
	ŋ	arngar	sing
Liquids and glides	l	arlar	large
	r	ara	run
	w	arwar	wow
	j	aryar	yard

Table 4.5: Prompts for consonants in the AVOZES data corpus. Phonemes marked with an asterisk (*) were omitted from the recordings.

4.5 Module 5 — Application Sequences - Digits

The sequences in this module can be used as examples of applying any results, gained from an analysis of the phonemes and visemes in the "short words" module, to short sequences that are more application-driven. Digit recognition is a common task in automatic speech recognition, e.g. [14, 30, 33], and similar sequences can be found in a number of AV speech corpora, for example in DAVID [3] and Tulips1 [22].

The AVOZES data corpus includes one sequence per digit for each speaker, spoken in order from 0 to 9. Again, 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.

4.6 Module 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/

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.

[5]This sentence appeared first in the corpora M2VTS and XM2VTSDB [19].

Chapter 5

Recording Setup

5.1 Recording Studio Layout

Recordings of the AVOZES data corpus were made in August 2000 and August 2001. The recordings took place in the audio laboratory of the Computer Sciences Laboratory (CSL) at the Australian National University. The same equipment was used on both occasions. The CSL audio 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 noise from the room's air-conditioning and the recording equipment. The computer and DV recorder were housed in an acoustic insulated box to reduce the amount of acoustic noise produced by the hard disk and cooling fans.

Figure 5.1 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. This light source was a normal office desk lamp with a reflective lampshade. Placing the light source below the cameras ensured a well-lit face, while blinding was reduced to a minimum. Other light source arrangements were considered, such as putting one light source on either side of the cameras — sufficiently apart so as not to blind the speaker (similar to [30]) — or a more expensive lighting system, as used in professional photographic studios. However, these were discarded in favour of the simplicity of a single light source, which achieved the objective of removing shadows in the mouth region. In addition, there was a general illumination of the room from three ceiling lights (normal light bulbs, not fluorescent light).

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

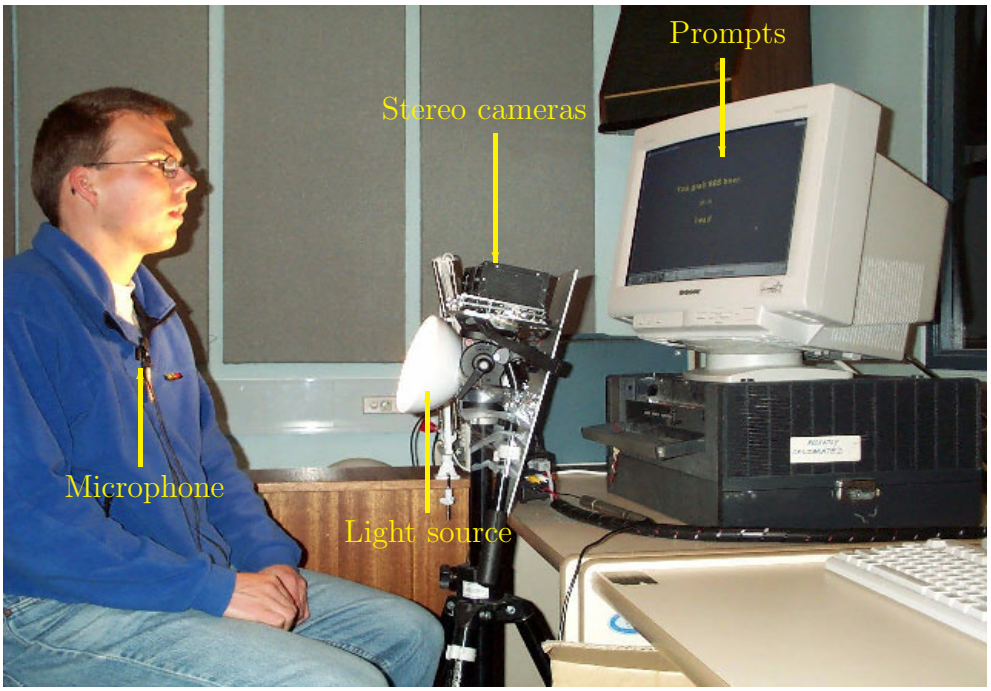

Figure 5.1: Recording setup in the CSL audio laboratory.

the camera arrangement etc. unchanged. Secondly, the process of building a face model for each speaker required sequences, in which the speaker was asked to turn the head 45° to the left and to the right of the cameras (see Section 4.2). The speakers were instructed to turn not just their eyes, but the whole head, so that it would point to corresponding markers on the wall. By sitting on a swivel chair, the speakers could, in fact, simply turn their whole bodies towards the markers. Keeping the vertical axis of the chair at a marked position ensured that the face was kept in the cameras' viewfields. The distance from the face to the cameras was about 600 ± 50mm, which corresponded to the distance ("depth") range that the cameras were calibrated for (see Section 5.4). 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.

5.2 Prompts

The speaking prompts appeared on the computer screen above the cameras. Figure 5.2 shows the stereo cameras in the foreground and the computer screen in the background from the viewpoint of a speaker. The distance

Figure 5.2: Speaker's view of the recording setup and the prompts on screen.

from the face of a speaker to the cameras varied between 550–650mm. The computer screen, from which the prompts were read, was another 20cm (in horizontal direction) behind the cameras (Figure 5.1). 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, so as to give the speaker an additional visual signal that a new prompt had appeared on the screen.

5.3 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 [34]. The microphone system was directly connected to the DV recorder, where the microphone's output was recorded as mono sound on DV tape with a 48kHz

sampling frequency.[1] The DV recorder was a JVC HR-DVS1U miniDV/S-VHS video recorder, which also featured an IEEE-1394 DV in/out connector.

To be able to perform various and repeated experiments on the same material of a speech data corpus, the sequences must be stored on a medium that allows easy repeated access, without loss of quality. For small corpora with few speakers and sequences, storage on a computer hard disk is possible. However, for large corpora with many speakers and sequences, such storage becomes quickly impossible or very expensive, despite ever-growing hard disk capacities. Video and audio compression is a way of overcoming problems with large amounts of data to some extent, but high compression is often related to loss of detail, which is clearly not desirable. Digital Video (DV) systems offer a good alternative for high-quality storage (see Appendix A for more details on the DV standard). MiniDV tapes are inexpensive and common tape sizes can store up to 63 minutes of video and audio data.

The two video cameras are 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 [17], then sent to a Hitachi IP5005 video card. In this technique, a device containing a video switching integrated circuit selects the signal from one video stream as the odd field of the video output, while the signal from the other video stream becomes the even field. This requires to first de-interleave the odd-even fields of the video frames from each camera. Multiplexing video signals in the analogue phase has the advantage that it can be applied to virtually any video hardware system. Images from two cameras can be stored in a single video frame. Stereo image processing can be performed within the computer's memory using only one image processing board. Single video stream processing is thus transformed into stereo vision processing.

A weakness of the field multiplexing technique is that only half the vertical resolution of the original video frame from each camera is available, as two video streams are compressed into a single frame. However, this disadvantage is more than outweighed by the ability to perform stereo vision processing with a single video card. It would be worthwhile in future studies to 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 of 16.6ms between the images from the two video streams, which is inherent in the NTSC standard, or any other interlaced video/TV standard.

[1] In the 48kHz sampling mode, two channels are recorded for stereo audio but in case of mono audio input, both channels contain the same signal.

That is, first all the lines of one field, let's say the odd lines, are processed, then all the lines of the other field. The field frequency is 60Hz in the NTSC standard, or 30Hz frame frequency, and hence there is a 16.6ms delay between fields. In the author's experience, this delay has not posed a problem in the analysis of the data, but it is important to be aware of this potential error source.

As already mentioned, the video stream lags the audio stream by one video frame, due to the recording setup. This delay must be taken care of in any analysis of the data.

In the Hitachi IP5005 video card, the video signal was unscrambled, so that the video sequences on tape show the output from the left camera in the top half and the output of the right camera in the bottom half of each video frame (as shown in Figure 5.3). The video signal was then sent from the video card to the DV recorder, where it was recorded as an NTSC YUV 4:1:1 signal at 29.97Hz frame rate (see Appendix A for an explanation of these terms). Because of the way that the outputs from the two cameras were multiplexed, there is a 16.6ms delay between the output from each camera in any recorded video frame. While virtually undetectable by the human eye at normal video play rate, it is a potential error source for the 3D reconstruction process, which requires the same object point to be identified in both images (and assumes that the object has the same shape in both the left and right image).

5.4 Camera Calibration

Camera calibration is the process of relating the camera's image (pixel) coordinates to the world coordinates. The relationship between the coordinate systems is described in the perspective transformation matrix. In the most general case, neither the intrinsic nor the extrinsic camera parameters are known. Intrinsic parameters define the perspective transformation from 3D object coordinates in the camera world coordinate system to the 2D camera image coordinate system. These parameters are

- f: focal length (or distance from image plane to centre of projection),
- κ_1, κ_2: lens distortion coefficients for both directions in image plane,
- s_x: uncertainty scale factor due to camera scanning and acquisition timing error,
- (u_O, v_O): coordinates of origin of image coordinate system in image plane.

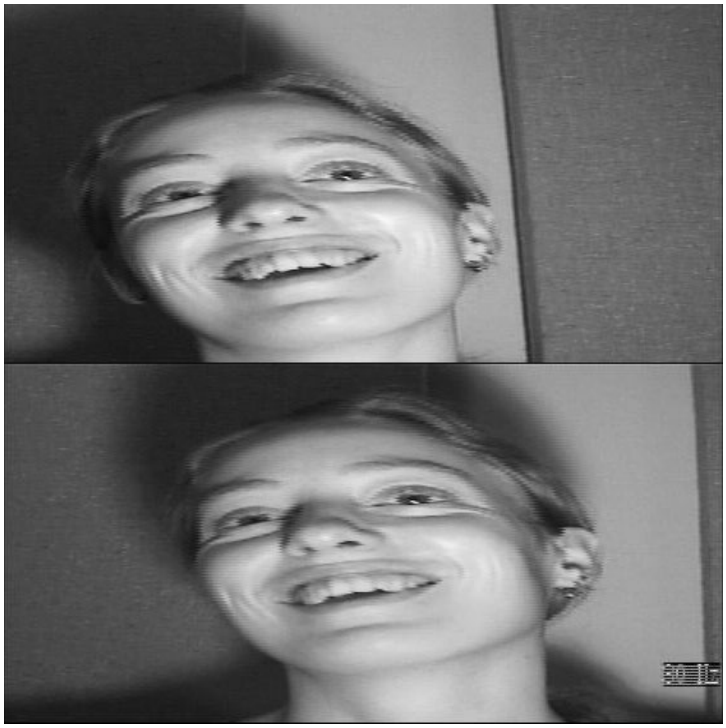

Figure 5.3: Example of stereo video put into one video frame. Left camera output in top half, right camera output in lower half.

Extrinsic parameters define the transformation from the 3D object world coordinate system to the 3D camera world coordinate system. In detail, these parameters are

- γ, ϑ, ϕ: rotation angles,
- $T = (t_x, t_y, t_z)^T$: elements of the translation vector.

5.4.1 Calibration Methods

Tsai [36] developed a camera calibration technique, for both a single camera system as well as stereo camera systems, that takes all of these 12 camera parameters into account. It is common to not calibrate the camera(s) for some parameters to simplify (and speed up) the calibration process by reducing the number of corresponding image points required. For example, if a perfect linear perspective transformation and no lens distortion are assumed, then the intrinsic parameters κ_1 and κ_2 can be omitted. Faugeras and Toscani [4] presented another approach to the calibration problem in stereo camera systems that assumes such a perfect perspective transformation.

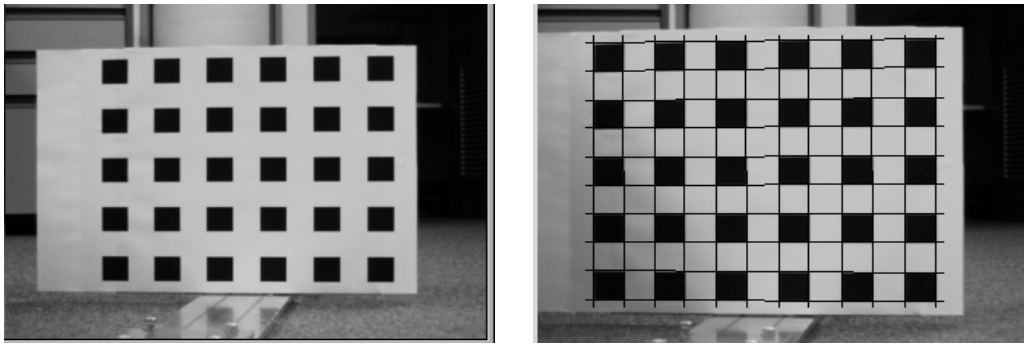

Figure 5.4: The calibration pattern: normal (left) and after edge detection (right).

Newman's Two-Step Process

However, the approach taken by Newman [25] in the RSL face tracking system is slightly different in that each camera is calibrated separately but using the same algorithm. As mentioned in the previous paragraph, a linear perspective transformation is assumed and non-linear camera effects (lens distortion) are not considered. Camera calibration is achieved in a 2-step process

1. Define a set of known 3D points in the scene and determine their image coordinates in the image plane.

2. Determine the perspective transformation matrix which maps the 3D object points onto their 2D image points.

In the first step, the stereo camera rig is placed on one end of an exactly measured calibration rig. The cameras observe an object plane parallel to the plane defined by the x and y coordinate axes. It features a rectangular 5×6 grid of 30 black rectangles on a white background similar to the grid used by Tsai [36] (Figure 5.4). The object plane is placed at various distances from the stereo camera rig, which are known exactly from the process of manufacturing the calibration rig.

The four corners of each rectangle are semi-automatically (the user has to click the mouse pointer on the corner rectangles to start the process) detected using edge detection in snapshots from both cameras. This procedure is repeated for all five positions (650–850mm) in which the object plane is placed. In total, this gives 30 rectangles × 4 corners × 5 positions = 600 corresponding image points for determining the 10 intrinsic and extrinsic camera parameters. The procedure takes only a few minutes and can be

done offline, before using the stereo camera system for face tracking or any other application.

In the second step, a minimisation procedure is usually necessary because of errors introduced by image noise and incorrectly located corresponding image points. As described in Section 6, the error between the measured and predicted 2D positions is minimised. Many papers in the literature describe general non-linear minimisation techniques (see [6, 36] for good overviews). Instead, a direct method, proposed in [6] and described in [35], is used here, because it is more accurate. Here, the perspective transformation matrix is determined by finding a matrix A such that for all i

$$A = \begin{pmatrix} a_1^T \\ a_2^T \\ a_3^T \end{pmatrix} \qquad u_i = \frac{a_1^T \cdot \vec{m}_i}{a_3^T \cdot \vec{m}_i} \qquad v_i = \frac{a_2^T \cdot \vec{m}_i}{a_3^T \cdot \vec{m}_i} \tag{5.1}$$

where $(u_i,\, v_i)$ are the image coordinates of the i^{th} calibration point with world coordinates $\vec{m}_i = (x_i,\, y_i,\, z_i)^T$. Errors in the measured image points $(u_i,\, v_i)$ make it practically impossible to satisfy these equations exactly, so A is found by minimising

$$E \;=\; \sum_i \Big((a_3^T \cdot \vec{m}_i)\, u_i - a_1^T \cdot \vec{m}_i \Big)^2 \;+\; \Big((a_3^T \cdot \vec{m}_i)\, v_i - a_2^T \cdot \vec{m}_i \Big)^2 \quad . \tag{5.2}$$

The camera parameters, and hence the transformation matrix, can then be extracted from the elements of A. However, because of noise, the rotational parameters will not necessarily form an exact rotation matrix. Choosing the closest rotation matrix may not minimise the error E any more. Instead of employing an iterative non-linear minimisation procedure, a more precise algorithm was developed by Newman *et al.* [26]. It can be shown that along each ordinate representing a rotation angle, the differential of E is quartic in that angle's cosine. Similarly, E is quadratic along the ordinates of all other camera parameters. Since closed form solutions of any quadratic, as well as the roots of any quartic, can be obtained, E can be minimised precisely. A small number of iterations (≈ 5) is sufficient to find the minimum. The resulting calibration is accurate to within 1mm over the range ($x = \pm 200$mm, $y = \pm 200$mm, $z = 600 \pm 300$mm) [27].

5.4.2 Results of Camera Calibration

The results of the camera calibration process for the sequences provided in the AVOZES data corpus are shown in Table 5.1. The calibration results for the focal length of each camera are given in millimeters. The image centre

Parameter	Camera	Coordinate	Calibration
Focal length	Left	X	1264.4118854
		Y	636.9071108
	Right	X	1259.8590426
		Y	634.9837795
Image centre	Left	X	238.3811862
		Y	122.5106600
	Right	X	267.6287652
		Y	114.6237397
Camera centre	Left	X	49.8668342
		Y	-3.8738391
		Z	-11.5996068
	Right	X	-59.4713793
		Y	-2.4802205
		Z	-16.9311634

Parameter	Camera	Calibration		
Rotation matrix	Left	0.9968766	0.0028878	-0.0789216
		-0.0055442	0.9999728	-0.0048641
		0.0789056	0.0052865	0.9968681
	Right	0.9952867	-0.0026490	0.0969404
		-0.0008564	0.9999127	0.0131875
		-0.0969671	-0.0132084	0.9951999

Table 5.1: Results of the camera calibration process for the sequences recorded in the AVOZES data corpus.

coordinates are in the local image coordinate system of each camera (see Chapter 6). The camera centre coordinates are in the stereo world coordinate system, which has its origin located roughly between the two cameras, as shown in Figure 6.1. The layout of the rotation matrices corresponds to that of other rotation matrices typically used in computer vision applications.

5.4.3 Discussion of Error Sources in Camera Calibration

Two main error sources can be identified for the calibration process. Firstly, camera lenses can show some radial and tangential distortions, which can be accounted for by the lens distortion coefficients κ_1 and κ_2. As a result of such lens distortions, the epipolar constraint may not hold. However, a perfect

linear transformation is often assumed and the effects of lens distortions are neglected, because the effects are considered small and omitting the determination of κ_1 and κ_2 simplifies the calibration process. Secondly, errors occur during the determination of corresponding points in the stereo images. This can be due to inaccuracies in an automatic determination (depending on the method(s) used), incorrectly manually chosen points, image quantisation, the delay between left and right images in the stereo vision system used in this project, and the fact that the cameras view the scene from different angles. The last point presents no problem for salient image features (the corner of a cube, for example), but may lead to incorrectly chosen correspondences for points on smooth surfaces. By using a specific calibration pattern of exactly known dimensions, as was done for the camera calibration before the recording of AVOZES, the problems of finding corresponding points in the stereo images can be avoided or at least reduced to a negligible level.

Chapter 6

From 2D to 3D - Stereo Reconstruction

The standard pinhole camera model can be assumed here, because any non-linear camera effects (radial and tangential lens distortions) are relatively small compared to the errors due to noise and stereo matching inaccuracies. In this model, the camera performs a linear perspective projection of an object point onto a pixel in the image plane through the camera centre. The camera arrangement and world coordinate system are shown in Figure 6.1. The cameras' centres, $\vec{c}_l$ and $\vec{c}_r$, are located equidistantly (about 55mm in the experiments) from the origin of the world coordinate system on the x axis. The y axis is vertically upwards and the z axis points horizontally out into the scene.

It is important to understand and distinguish the various coordinate systems that will be referred to in the following. First of all, there is the *image coordinate system of each camera*. This is a 2D coordinate system, which is represented by (u, v) coordinates for the left camera and (r, s) coordinates for the right camera, respectively. Secondly, there is the *world coordinate system of each camera*. These are 3D coordinate systems with the origin (= centre of projection) in the camera centre. Finally, there is the *stereo world coordinate system*, depicted in Figure 6.1, with its origin halfway between the two camera centres.

6.1 Epipolar Geometry

Figure 6.2 shows the *epipolar geometry*, which is the basic constraint arising from having two cameras (at different locations = viewpoints) looking at the same scene. A very good introduction into epipolar geometry can be found

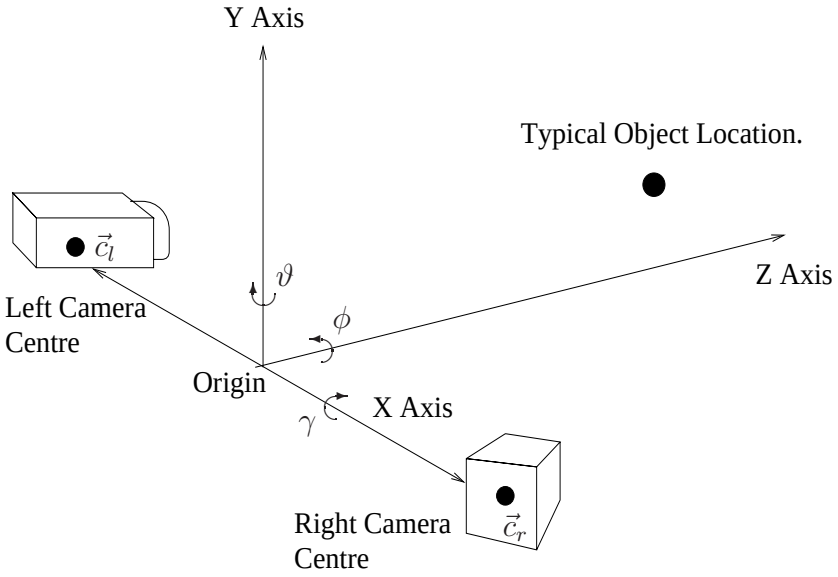

Figure 6.1: Stereo camera arrangement and stereo world coordinate system.

in [38]. The line through the two camera centres, $\vec{c}_l$ and $\vec{c}_r$, projects to a point $\vec{e}_l$ in the left image plane and $\vec{e}_r$ in the right image plane. The points $\vec{e}_l$ and $\vec{e}_r$ are called *epipoles*. The camera centres $\vec{c}_l$, $\vec{c}_r$ and point $\vec{m}$ form a plane — the *epipolar plane* for the point $\vec{m}$. The image points, $\vec{m}_l$ and $\vec{m}_r$, must lie on the *epipolar lines* l_{m_l} and l_{m_r}, respectively. These epipolar lines are defined by the intersection of the epipolar plane with the image planes of the cameras and must therefore, by definition, go through the epipoles.

An algorithm for computing the 3D structure of a scene from a pair of perspective projections, where the spatial relationship between the two views is unknown, was first presented by Longuet-Higgins [13]. He showed that if a scene contains at least eight corresponding points in the images from both views, the relative orientation of the two projections and the structure of the scene can be computed by solving a set of simultaneous linear equations based on the eight sets of image coordinates. This only accounts for extrinsic camera parameters, i.e. rotation and translation (see Section 5.4 for an explanation of camera parameters). The relationship between corresponding image points in the two camera images is described in the *Essential matrix* $\mathbf{E}$ — a 3×3 matrix — and satisfies

$$\vec{m}_r^T \, \mathbf{E} \, \vec{m}_l = 0 \quad . \tag{6.1}$$

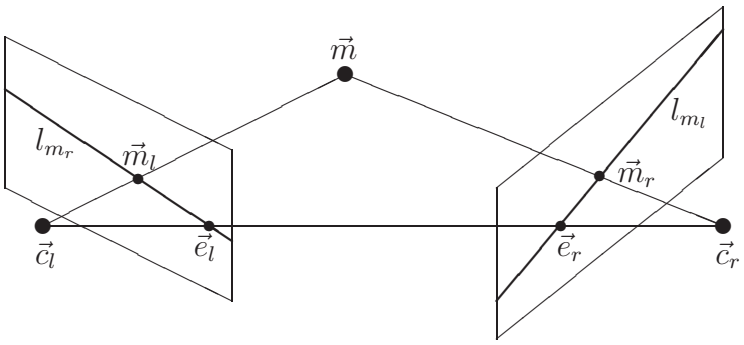

Figure 6.2: Epipolar Geometry.

Luong and Faugeras [16] generalised Longuet-Higgins' algorithm to also include intrinsic camera parameters (see Section 5.4). The relationship between corresponding image points is expressed in the 3×3 *Fundamental matrix* $\mathbf{F}$, which can be computed from coordinates of corresponding points in uncalibrated images, see [15, 16] for details. The Fundamental matrix satisfies

$$\vec{m}_r^T \, \mathbf{F} \, \vec{m}_l = 0 \quad . \tag{6.2}$$

Let us denote the (2D) image point of an object point $\vec{m}_i$ in the left and right image planes respectively by

$$\vec{m}_i^l = \begin{pmatrix} x_i^l \\ y_i^l \\ z_i^l \end{pmatrix} \quad \text{and} \quad \vec{m}_i^r = \begin{pmatrix} x_i^r \\ y_i^r \\ z_i^r \end{pmatrix} \tag{6.3}$$

with the z_i element representing the distance of the image plane from the camera centre. It is generally more convenient to use homogeneous coordinates, which can be established by dividing the vector elements by the element in the third row

$$u_i = \frac{x_i^l}{z_i^l}, \quad v_i = \frac{y_i^l}{z_i^l}, \quad r_i = \frac{x_i^r}{z_i^r}, \quad s_i = \frac{y_i^r}{z_i^r}, \tag{6.4}$$

$$\vec{m}_i^l = \begin{pmatrix} u_i \\ v_i \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{m}_i^r = \begin{pmatrix} r_i \\ s_i \\ 1 \end{pmatrix} \quad . \tag{6.5}$$

(Almost every textbook on computer graphics or computer vision will discuss the use of homogeneous coordinates in detail, for example, consult [5].)

The *perspective transformation matrix* (or *camera calibration matrix*) defines the transformation from image coordinates to camera world coordinates.

It is determined during camera calibration as described in Section 5.4. If matrices for both cameras are known, $\vec{m}_i^l$ and $\vec{m}_i^r$ can be transformed into vectors in camera world coordinates. Not considering non-linear camera effects, this matrix represents a rotation as well as a translation. If the centre of projection coincides with the camera centre (and origin of each camera's world coordinate system), then the translational component equals 0.

The resulting vectors $\vec{p}_i$ and $\vec{q}_i$ represent directions from the camera centres, through the respective point on the image plane, to the object point in the scene

$$\vec{p}_i = R_y(\vartheta_l)\, R_z(\phi_l)\, R_x(\gamma_l)\, f_l\, \vec{m}_i^l \tag{6.6}$$
$$\vec{q}_i = R_y(\vartheta_r)\, R_z(\phi_r)\, R_x(\gamma_r)\, f_r\, \vec{m}_i^r \tag{6.7}$$

The scalars f_l and f_r are the focal lengths of the left and right cameras, respectively. R_x, R_y, and R_z are rotations around the x, y, and z axes, respectively. As mentioned in the previous subsection, the cameras in this project were mounted on a rig with a baseplate in the xz plane which allows one to verge the cameras around the y axis, but limits rotation around the other two axes. In this study, the angles were $\vartheta_l \approx -5°$, $\vartheta_r \approx 5°$, and $\phi_l \approx \phi_r \approx \gamma_l \approx \gamma_r \approx 0$.

Under ideal conditions, the vectors $\vec{p}_i$ and $\vec{q}_i$ intersect at the 3D point $\vec{m}_i = (x, y, z)^T$. However, since $\vec{p}_i$ and $\vec{q}_i$ are likely to be corrupted by noise (lens distortion, point correspondence), $\vec{m}_i$ is determined by minimising the error term

$$E_i = \|\vec{p}_i\, s_i + \vec{c}_l - \vec{m}_i\|^2 + \|\vec{q}_i\, t_i + \vec{c}_r - \vec{m}_i\|^2 \tag{6.8}$$

with respect to the three coordinates of $\vec{m}_i$ and the two scalars s_i and t_i. If stereo matching fails, i.e. if the image points $\vec{m}_i$ are found incorrectly, minimising the error term E_i will not determine the coordinates of $\vec{m}_i$ correctly. Finding matching image points — solving the 'correspondence problem' — is therefore of great importance.

The final step in 3D reconstruction is the stereo triangulation, which leads to the coordinates of $\vec{m}_i$ in stereo world coordinates. If the orientation and distance of each camera to the origin of the stereo world coordinate system is known, then, together with the perspective transformations of each camera, the relative orientation of the two cameras to each other and the stereo world coordinates of a point viewed in both camera images can be calculated.

Setting the partial derivatives of E_i to zero gives the following solution for $\vec{m}_i$

$$P_i = \frac{\vec{p}_i \cdot \vec{p}_i^T}{\|\vec{p}_i\|^2} - \mathbf{I} \qquad Q_i = \frac{\vec{q}_i \cdot \vec{q}_i^T}{\|\vec{q}_i\|^2} - \mathbf{I} \tag{6.9}$$

$$(P_i \; + \; Q_i)\, \vec{m}_i \;=\; P_i\, \vec{c}_l + Q_i\, \vec{c}_r \quad . \tag{6.10}$$

Inverting the matrix coefficient $(P_i + Q_i)$ yields the three coordinates of $\vec{m}_i$.

Chapter 7

Recorded Data

7.1 Sequences

AVOZES currently contains recordings made from 20 native speakers of AuE. The group is gender balanced with ten female and ten male speakers. Six speakers wear glasses, three wear lip make-up, two have beards. Figure 7.1 shows the faces of the native speakers of AuE. 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 the accent varieties, it is important to point out that the 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 speech varieties are not discrete entities, but rather span a continuum of accent variation, so that any classification can only be approximate.

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. In addition to these 20 native speakers of AuE, four speakers with a different language background were also recorded. At the first recording, sequences from three such speakers were taken. Two of them have an English language background (United Kingdom, New Zealand), the third speaker speaks German as his first language, but had spent one year in the United Kingdom and two years in Australia at the time of recording. At

Figure 7.1: Face shots of the 20 native speakers of AuE in AVOZES.

the second recording, one non-native English speaker with a Chinese dialect as his first language, but who had lived in Australia for 6 years at the time of recording, was recorded. The non-native speaker data is currently not published, but may be published in future.

Each speaker spent about half an hour in the recording studio. They were first familiarised with the speech material and informed about the recording procedure about to follow. Actual recordings took about five minutes per speaker. A total of 56 sequences were recorded per speaker (1 face sequence and 55 speech material sequences). The author was present as a record-

ing assistant, so that speakers did not have to handle any of the equipment themselves and could concentrate on the speaking task. All sequences are available as AVI files containing both audio and video information as well as WAV audio files only containing the audio component. They were produced from the sequences on the DV tapes using MGI VideoWave 4. Video information is encoded using the NTSC format, 720×480 pixels, 29.97Hz frame rate. The AVOZES AVI files use the Adaptec DVSoft codec, which most media players like RealPlayer, Windows Media Player, etc. have pre-installed. Audio information is encoded as 48kHz, 16-bit stereo (although only a mono microphone was used, i.e. the two stereo channels contain the same information). The length of the individual sequences has typically been chosen to be a multiple of a full second (i.e. of 30 video frames).

It shall be mentioned again that the video infoirmation lags the audio information by one video frame in the AVOZES AVI-files, due to the recording setup using field multiplexing and descrambling before the video stream was recorded on tape. The audio stream was recorded directly from the microphone onto the DV tape without a delay due to processing. The AVOZES AVI-files contain the data like it was recorded on tape. No correction of the delay has been performed.

The AVOZES data corpus currently contains only frontal face ($\pm 10^{\circ}$) AV speech recordings, with no separate or simultaneous recordings from a different angle. The faces were illuminated from the front. Recordings were made for a clean audio condition. There was no particular background noise other than what has already been described in Section 5.1. However, artificial (computer-generated) noise could be added to the audio signal, if that was desired for some experiment. In that way, the control of the noise is much better, because it can be designed to suit a particular experimental situation and the AVOZES date corpus could be used for a wider range of tests. It should be noted that adding artificial noise does not take account of the Lombard effect [12]. Recordings were made in a conversational tone.

7.1.1 Complete Recordings

AVOZES comes normally as one file per sequence, i.e. there are 56 AVI- and 56 WAV-files per speaker. However, the complete, continuous recordings, from which the individual sequences were edited, are also available upon request. In that case, only one AVI- and one WAV-file per speaker can be found on the DVDs. These raw sequences contain additional video frames and audio samples from times between individual sequences. They may be useful to some researchers.

7.2 Speaker Data

Beside the actual recordings, each speaker was also asked to fill in a form about personal data, so that any outstanding effects in the recorded material could be checked against these data. A similar approach was taken in the ANDOSL database [20]. It is important to collect such data in addition to the signal data, as for example professional training in singing or medical conditions of the respiratory system can have an effect on the pronunciation. Personal data collected contains:

- name, 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 range was 150–250mm. The individual information (names omitted, date of birth transformed into age) about the native speakers of AuE is presented in Appendix B.

7.3 Sequence File Names

This section provides a guide to what you can find on the DVDs. The top level directory contains the general files, showing the scene without any speakers (Module 1), as well as the top level of the subdirectories for each of the 20 native speakers of AuE in AVOZES. The latter are labelled `f1-f10` for the female speakers and `m1-m10` for the male speakers, and correspond to the speaker labels used in Appendix B and elsewhere in this document. All of the speaker-dependent sequences have the same file name for all speakers

except for a prefix reflecting the speaker label, e.g. the general face sequences are named `f1Face.avi`, `f2Face.avi`, `f3Face.avi` and so on.[1]

On the next directory level, each speaker directory contains five subdirectories named

- `RecordingSetupWithSpeaker` (Module 2),
- `CalibrationSequences` (Module 3),
- `Phonemes` (Module 4),
- `Digits` (Module 5), and
- `ContinuousSpeech` (Module 6).

Module 2 does not have any further subdirectories and only contains the file `m1Face.avi`. Module 3 has two subdirectories named `Bababa` and `Ioioio` containing the corresponding files `m1Bababa.avi` and `m1Ioioio.avi`, respectively.

Module 4 contains the core sequences of AVOZES, which cover all visemes and almost all phonemes of AuE (see Section 4.4). The directory `Phonemes` contains two subdirectories: one for the vocalic phonemes (`CVCWords`) and one for the consonantal phonemes (`VCVWords`). These directories then contain further subdirectories named after the prompts listed in Tables 4.4 and 4.5, e.g. `Bab`, `Babe`, `Barb`, and so on, which in turn then contain the correspondingly named AVI- and WAV-files, e.g. `m1Bab.avi`, `m1Babe.avi`, `m1Barb.avi`, and so on.

Modules 5 and 6 contain sequences that are more application-driven. Module 5 has ten subdirectories, one each for the digits from 0 to 9. Each of these directories then contains the corresponding files, e.g. `m1Zero.avi`, `m1One.avi`, `m1Two.avi`, and so on. Module 6 has three subdirectories named after the beginning of the continuously spoken sentences listed in Section 4.6, i.e. `JoeTook`, `ThinHair`, `YesterdayMorning`. These directories then contain the correspondingly named AVI- and WAV-files.

Figures 7.2 and 7.3 give an overview of the directory structure as found on the DVDs.

[1] In the following, only the file names for one speaker and with `.avi` ending are listed. It can be automatically assumed that the corresponding `.wav` file is present in the same directory and that the same structure and file names exist in the appropriate directories of the other speakers.

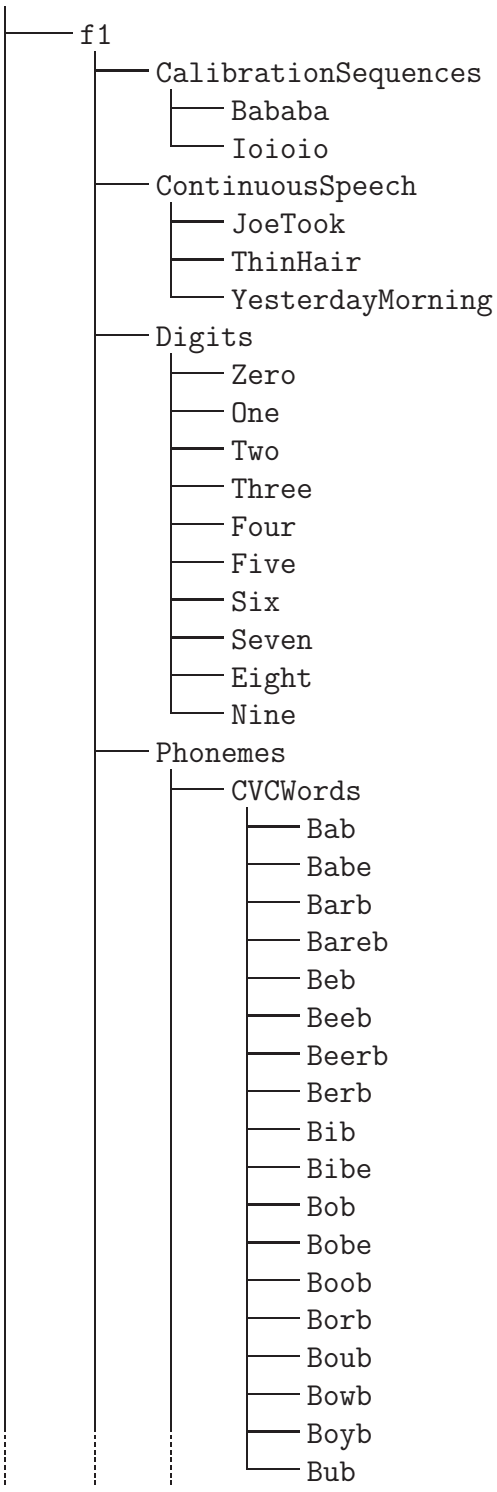

Figure 7.2: The directory structure of the AVOZES sequences as provided on the DVDs (Part 1).

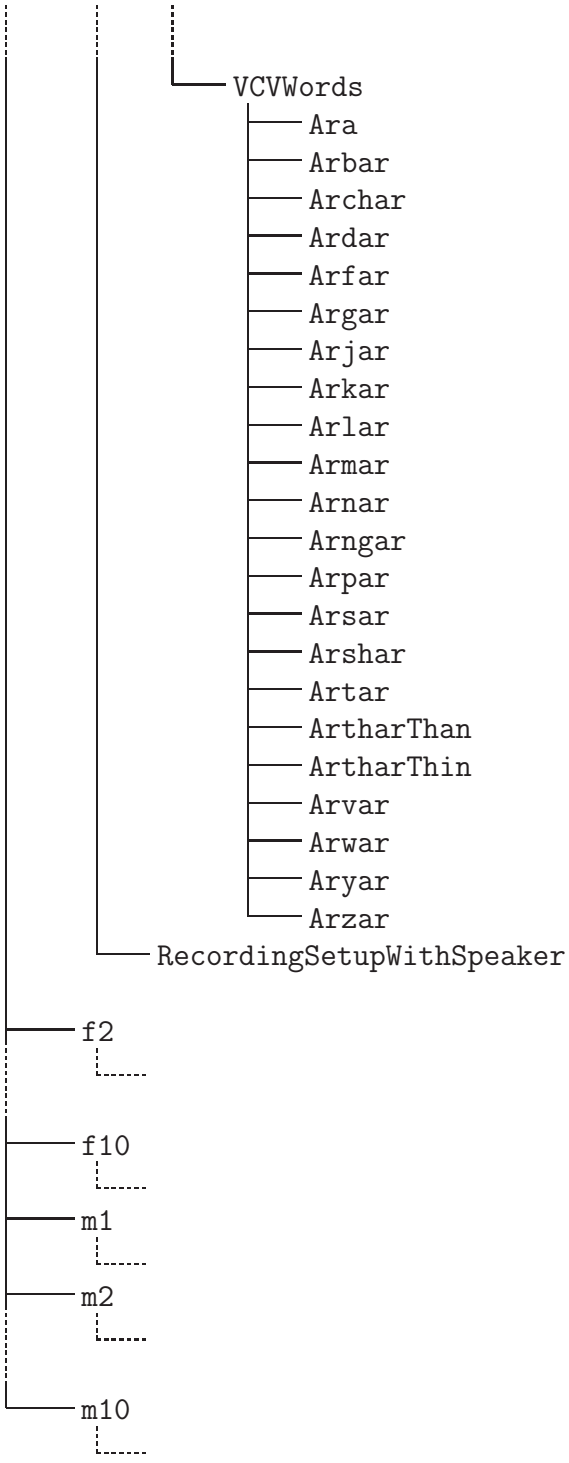

Figure 7.3: The directory structure of the AVOZES sequences as provided on the DVDs (Part 2).

Chapter 8

Allowed Usage of the AVOZES Data Corpus

Using the data of the AVOZES data corpus requires a licence. If you haven't got a licence yet, but would like to use AVOZES, please contact the author (see title page for contact details). A licence can be acquired by individuals, institutions, or commercial entities. Most users are anticipated to come from an academic institution and would therefore be interested in a non-commercial (academic) licence. The rest of this chapter describes the allowed use under such a non-commercial licence. If you are interested in a commercial licence, please contact the author.

8.1 Non-Commercial (Academic) Licence

This licence is primarily intended for users at academic or other non-commercial research institutions for the purpose of evaluation and internal research. It is most likely of particular use for research in AVSP, especially on AuE, but may also be useful for research in other fields. However, the data is provided "as is" and no warranty is given that it is useful or appropriate for the user's research. For the complete wording of this licence, please contact the author.

Under the non-commercial licence, the user is allowed to make as many copies of the data as is reasonably necessary for back-up purposes and use. The user is not allowed to alter or modify the data in any way, unless prior written permission has been given by the licensor. The user is also not allowed to combine the data with or incorporate the data in any other data for the purpose of publication or external use of such data, unless prior written permission has been given. However, the user is explicitly granted

the right to perform experiments with and analyses on the data, use the data in conjunction with other (self-recorded or otherwise acquired) data, and may publish the results of any such work, under the obligation to explicitly make a reference to using AVOZES and cite the following paper:

R. Goecke and J.B. Millar "The Audio-Video Australian English Speech Data Corpus AVOZES", *8th International Conference on Spoken Language Processing ICSLP2004*, Jeju Island, Korea, 4-8 October 2004.

Furthermore, the user is allowed to extract video frames ("frame grabbing") and audio samples for the purpose of including them in the user's research publications (journal papers, conference papers, student theses) and presentations (conferences, lectures, seminars, web pages), provided that no such publication or presentation contains more than 50 video (still) frames and / or more than 10 audio or AV sequences[1], unless prior written permission has been given by the licensor, and provided that the AVOZES data corpus is referenced as described above. If in doubt, the user should contact the author or licensor first.

If the results of any work using the AVOZES data corpus leads to the commercialisation of the results, and any data of the AVOZES data corpus becomes part of the commercialised product or products derived from it, a commercial licence for AVOZES must be acquired.

[1]Where sequence refers to the AVI- and WAV-files provided on the DVDs.

Appendix A

Digital Video Format

A short overview of the Digital Video (DV) format is given here for the interested reader. Detailed information can be found in the international standards document IEC 61834 [10].[1] The original DV format (or Digital Video Cassette (DVC)) standards document is the so-called "Blue Book" [1]. The DV format should not be confused with standards for DVD (Digital Video Disc or Digital Versatile Disc) or DVB (Digital Video Broadcasting), which are different.

DV is an international standard for a consumer digital video format created by a consortium of ten companies. The companies originally involved in creating the standard were Matsushita Electric Industrial Corp (Panasonic), Sony Corp, Victor Corporation of Japan (JVC), Philips Electronics N.V., Sanyo Electric Co. Ltd., Hitachi Ltd., Sharp Corporation, Thompson Multimedia, Mitsubishi Electric Corporation, and Toshiba Corporation. Since then others have joined; there are now over 60 companies in the DV consortium.

The sampled video is compressed using a Discrete Cosine Transform (DCT), the same sort of compression used in motion-JPEG and MPEG. However, DV's DCT allows for more local optimisation (of quantising tables) within the video frame than do JPEG compressors, thus allowing for higher quality at the nominal 5:1 compression factor than a JPEG frame would show. DV uses intraframe compression; each compressed frame depends entirely on itself, and not on any data from preceding or following frames. However, it also uses adaptive interfield compression. If the compressor detects little difference between the two interlaced fields — the odd and even fields — of a frame, it will compress them together, freeing up some of the 'bit budget' to allow for higher overall quality. In theory, this means

[1] The information presented here is largely taken from the websites www.dvformat.com and www.adamwilt.com .

that static areas of images will be more accurately represented than areas with a lot of motion. In practice, this can sometimes be observed as a slight degree of 'blockiness' in the immediate vicinity of moving objects.

There are different colour sampling models for digital video depending on the original input (NTSC, PAL, etc.). The colour sampling used in the work presented in this thesis was NTSC YUV 4:1:1. The first number refers to the sampling rate of the luminance (Y), the other two numbers refer to that of the colour difference signals (U and V) relative to the first one. In a 4:1:1 system, the colour difference signals are sampled every fourth luminance sample. Other common sampling structures are 4:2:2 (D-1, D-5, etc.) and 4:2:0 (PAL).

In NTSC DV format, the resolution is 720×480 pixels. DV sampling is mostly said to be at exactly 30Hz frame rate (or 60Hz field rate)[2] but it is actually at a frame rate of 29.97Hz. To keep in synchronisation with the NTSC TV frame rate of exactly 30 frames per second, the video sequence of one second per minute — usually the first — DV contains only 28 frames! It is important to take this into account when analysing DV data, for example by interpolating the 28 frames to 30 frames.

Audio sampling in the DV standard is PCM (Pulse Code Modulation) at 48kHz with 16bit (2 channels), at 32kHz with 12bit (4 channels), or at 44.1kHz with 16bit (2 channels, same as audio CD (CD-DA) sampling). PCM is a way to digitise analogue signals by sampling the signal at constant time intervals (e.g. [7]). The amplitude at each sampling point is rounded off to the nearest of several specific, predetermined levels in a process called quantisation. The error between the exact sample value and the assigned level is called quantisation error. The number of levels is defined by the number of bits assigned to represent each sample value, e.g. 16bit gives $2^{16} = 65,536$ levels. The 48kHz and 32kHz sampling rates of the DV standard can be used in locked mode, the 44.1kHz sampling rate only in unlocked mode. In locked mode, the audio sample clock is precisely locked to the video sample clock such that there is exactly the same number of audio samples recorded per video frame (or multiples of one video frame). To ensure synchronisation between the audio and video signals, locked mode is generally preferable.

Finally, the DV format is very well suited for transfer via an IEEE-1394 'FireWire' link to other equipment. For example, a computer with an IEEE-1394 compliant I/O-card can be linked to a DV recorder, so that there is a fully digital transfer of the data stored on DV tape to the computer for processing, analysis etc., thereby eliminating potential quality losses due to digital-to-analogue and analogue-to-digital conversion.

[2]Note, NTSC is an interlaced video standard.

Appendix B

Speaker Data

This appendix contains background information on the 20 native speakers of AuE in the AVOZES data corpus. The tables on the following pages summarise some background information on the speakers in the AVOZES data corpus. The speakers were asked to fill in a questionnaire at the time of recording and the answers are presented here. The kind of information collected is detailed below (Table B.1).

Speaker	Identifier, f1–f10 for female speakers,
m1–m10 for male	speakers
AuE variety	Variety of AuE: broad (B), general (G), cultivated (C)
Age	At the time of recording (in years)
Height	In cm
Weight	In kg
Level of education	Secondary, Tertiary, etc
Time abroad	Significant time spent abroad by the speaker
	(where, how long for, and at what age)
Singing / Training	Does the speaker sing? Has the speaker received
	training?
Smoking	Is or was the speaker a smoker?
Medical conditions	Related to respiratory system or otherwise
	potentially affecting the speech production
Microphone	Distance in cm between microphone and mouth
Country of origin	Of the speaker's parents
Native language	Of the speaker's parents
Occupation	Of the speaker's parents

Table B.1: Description of column headers in the tables on the following pages.

Speaker	AuE Variety	Age	Height	Weight	Level of education	Time abroad (where, how long)	Singing / Training	Smoking	Medical conditions	Microphone
f1	C	23	163	62	Tertiary	–	Yes / No	No	–	19
f2	G	47	168	58	Tertiary	Hungary 2yr age 0–2 Canada 4yr age 22–26	No / No	No	–	14
f3	G	23	169	72	Tertiary	–	No / No	No	–	23
f4	B	22	170	54	Tertiary	–	No / No	No	Fibromyalgia since age 19	12
f5	G	32	163	55	Tertiary	Malaysia 2yr age 21–23	Yes / Yes	No	–	12
f6	C	28	164	56	Tertiary	–	No / No	No	–	20
f7	G	38	172	63	Tertiary	USA 8yr age 29–37	No / No	Yes	–	16
f8	G	29	175	70	Tertiary	–	No / No	Yes	–	18
f9	G	24	164	50	Tertiary	–	No / No	No	–	17
f10	G	24	172	60	Tertiary	–	Yes / Yes	No	–	20

Table B.2: Speaker data for female speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Speaker	AuE Variety	Age	Height	Weight	Level of education	Time abroad (where, how long)	Singing / Training	Smoking	Medical conditions	Microphone
m1	G	40	178	90	Tertiary	Europe 6mo age 32 Japan 3yr age 33–35	No / No	No	–	24
m2	B	56	175	92	Tertiary	USA 1yr age 27	No / No	No	–	18
m3	B	27	188	95	Tertiary	USA 6mo age 8 UK 6mo age 8	No / No	No	Perforated ear drum, age 12	21
m4	B	26	175	62	Tertiary	–	No / No	No	No nasal breathing until age 13	20
m5	B	33	174	87	Tertiary	–	No / No	No	Asthma	16
m6	G	26	198	91	Tertiary	Scotland 1yr age 1 England 6mo age 13 USA 5mo age 15	No / No	No	–	26
m7	G	28	178	57	Tertiary	–	No / No	No	Mild asthma	25
m8	B	27	188	95	Tertiary	Argentina 8mo age 25	No / No	No	–	18
m9	G	27	190	75	Tertiary	NZ 15yr age 0–14 Japan 5mo age 25	No / No	No	–	18
m10	G	28	175	70	Tertiary	Switzerland 3mo age 17	Yes / No	No	–	14

Table B.3: Speaker data for male speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Speaker	Mother's			Father's		
	Country of Origin	Native Language	Occupation Occupation	Country of Origin	Native Language	Occupation
f1	Australia	English	Secretary	Australia	English	Retired
f2	Hungary	Hungarian	Education research officer	Hungary	Hungarian	Industrial chemist
f3	Australia	English	Consultant in education	Australia	English	Teacher
f4	Australia	German	Primary teacher	Australia	English	Furniture maker
f5	Australia	English	Housewife	Australia	English	Electrician
f6	Australia	English	Historian	Australia	English	Air Force Pilot
f7	Australia	English	Retired	Australia	English	Retired
f8	Australia	English	Hairdresser	Australia	English	Financier
f9	Australia	English	Caterer	Egypt	Greek	Architect
f10	Australia	English	Psychologist	Australia	English	Geologist

Table B.4: Family background for female speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Speaker	Mother's			Father's		
	Country of Origin	Native Language	Occupation Occupation	Country of Origin	Native Language	Occupation
m1	China	Russian	Housewife	Russia	Russian	Teacher
m2	Australia	English	Housewife	Australia	English	Plumber
m3	Australia	English	Librarian	New Zealand	English	Academic
m4	Australia	English	Physiotherapist	Australia	English	Storeman
m5	Australia	English	Retired	Australia	English	Retired
m6	Australia	English	Administrator	Australia	English	Mathematician
m7	Australia	English	Clinical nurse consultant	Australia	English	Courier
m8	Australia	English	Librarian	Australia	English	Engineer
m9	New Zealand	English	Accountant	New Zealand	English	Painter
m10	Australia	English	Manager of accounts	Australia	English	Medical doctor

Table B.5: Family background for male speakers. For a description of the column headers, see the beginning of this section. The native language of all speakers is English.

Bibliography

[1] Blue Book. *Specifications of Consumer-Use Digital VCRs using 6.3mm magnetic tape*. HD Digital VCR Conference, December 1994.

[2] C.C. Chibelushi, F. Deravi, and J.S. Mason. Survey of audio visual speech databases. Technical report, Department of Electrical and Electronic Engineering, University of Wales, Swansea, UK, 1996.

[3] C.C. Chibelushi, S. Gandon, J.S. Mason, F. Deravi, and D. Johnston. Design Issues for a Digital Integrated Audio-Visual Database. In *IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication*, pages 7/1–7/7, London, UK, Digest Reference Number 1996/213, November 1996.

[4] O.D. Faugeras and G. Toscani. The calibration problem for stereo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR'86*, pages 15–20, Miami Beach, USA, June 1986. IEEE.

[5] J.D. Foley, A. van Dam, S.K. Feiner, and J.F. Hughes. *Computer Graphics - Principles and Practice*. Addison-Wesley, Reading (MA), USA, 1996.

[6] S. Ganapathy. Decomposition of Transformation Matrices for Robot Vision. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 130–139, Atlanta (GA), USA, March 1984.

[7] J.D. Gibson. *Principles of digital and analog communications*. Macmillan, New York (NY), USA, 2nd edition, 1993.

[8] R. Goecke and J.B. Millar. The Audio-Video Australian English Speech Data Corpus AVOZES. In *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004*, Jeju, Korea, October 2004.

[9] R. Goecke, Q.N. Tran, J.B. Millar, A. Zelinsky, and J. Robert-Ribes. Validation of an Automatic Lip-Tracking Algorithm and Design of a Database for Audio-Video Speech Processing. In *Proceedings of the 8th Australian International Conference on Speech Science and Technology SST2000*, pages 92–97, Canberra, Australia, December 2000. Australian Speech Science and Technology Association (ASSTA).

[10] IEC. *61834*. International Electrotechnical Commission, consolidated edition, 2001.

[11] IPA. *Handbook of the International Phonetic Association*. Cambridge University Press, Cambridge, United Kingdom, 1999.

[12] J.C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93(1):637–642, 1993.

[13] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, September 1981.

[14] J. Luettin and S. Dupont. Continuous Audio-Visual Speech Recognition. In *Proceedings of the Fifth European Conference on Computer Vision ECCV'98*, volume II of *Lecture Notes in Computer Science LNCS 1407*, pages 657–673, Freiburg, Germany, June 1998. Springer-Verlag, Berlin, Germany.

[15] Q.-T. Luong, R. Deriche, O.D. Faugeras, and T. Papadopoulo. On Determining the Fundamental matrix: Analysis of Different Methods and Experimental Results. Technical Report 1894, Unité de Recherche INRIA-Sophia Antipolis, Institut National de Recherche en Informatique et en Automatique, Sophia-Antipolis, France, April 1993.

[16] Q.-T. Luong and O.D. Faugeras. The Fundamental matrix: theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17(1):43–76, 1996.

[17] Y. Matsumoto, T. Shibata, K. Sakai, M. Inaba, and H. Inoue. Real-Time Color Stereo Vision System for a Mobile Robot based on Field Multiplexing. In *Proceedings of the IEEE International Conference on Robotics and Automation ICRA'97*, pages 1934–1939, Albuquerque (NM), USA, April 1997. IEEE.

[18] K. Messer, J. Matas, and J. Kittler. Acquisition of a large database for biometric identity verification. In *Proceedings of BIOSIGNAL 98*, pages 70–72, Brno, Czech Republic, June 1998.

[19] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication AVBPA'99*, pages 72–77, Washington (DC), USA, March 1999.

[20] J.B. Millar, J.P. Vonwiller, J.M. Harrington, and P.J. Dermody. The Australian National Database Of Spoken Language. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing ICASSP'94*, volume 1, pages 97–100, Adelaide, Australia, 1994.

[21] J.B. Millar, M. Wagner, and R. Goecke. Aspects of Speaking-Face Data Corpus Design Methodology. In *Proceedings of the 8th International Conference on Spoken Language Processing ICSLP2004*, Jeju, Korea, oct 2004.

[22] J.R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858, Cambridge (MA), USA, 1995. MIT Press.

[23] J.R. Movellan and G. Chadderdon. Channel Separability in the Audio-Visual Integration of Speech: A Bayesian Approach. In D.G. Stork and M.E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series*, pages 473–487, Berlin, Germany, 1996. Springer-Verlag.

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

[25] R. Newman. Head Pose and Gaze Point Estimation System Version 2.0. Technical report, Robotic Systems Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, 1999.

[26] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. Real-Time Stereo Tracking for Head Pose and Gaze Estimation. In *Proceedings*

of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition FG'2000, pages 122–128, Grenoble, France, March 2000.

[27] R. Newman and A. Zelinsky. Error Analysis of Head Pose and Gaze Direction from Stereo Vision. In *Proceedings of the Australian Conference on Robotics and Automation ACRA'99*, pages 114–118, Brisbane, Australia, March 1999.

[28] T. Öhman. 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, 1998.

[29] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy. CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP2002*, volume 2, pages 2017–2020, Orlando (FL), USA, May 2002. IEEE.

[30] E.D. Petajan. *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, University of Illinois at Urbana-Champaign, 1984.

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

[32] G.L. Plant and J.J. Macrae. Visual Perception of Australian Consonants, Vowels and Diphthongs. *Australian Teacher of the Deaf*, 18:46–50, July 1977.

[33] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe. Speaker Independent Audio-Visual Database for Bimodal ASR. In C. Benoît and R. Campbell, editors, *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing AVSP'97*, pages 65–68, Rhodes, Greece, September 1997. ESCA.

[34] Sennheiser. *User's Guide: Handgrip/Powering Module K3N / K3U and Microphone Heads MKE 10-3 / ME 20 / ME 40 / ME 80 / ME 88*. Sennheiser Electronic KG, Wedemark, Germany, December 1981.

[35] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River (NJ), USA, 1998.

[36] R.Y. Tsai. An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR'86*, pages 364–374, Miami Beach (FL), USA, June 1986. IEEE.

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

[38] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion, and Object Recognition: A Unified Approach.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.