The Audio-Video Australian English Speech Data Corpus AVOZES

Roland Goecke[1,3] *and J Bruce Millar*[2,3]

[1]Fraunhofer IGD-R, Rostock, Germany, [2]Australian National University, Canberra, Australia,
[3] National ICT Australia[†], Canberra Laboratory

Corresponding author: roland.goecke@ieee.org

Abstract

This paper presents the Audio-Video Australian English Speech data corpus AVOZES. It contains recordings of 20 speakers uttering a variety of phrases. The corpus was designed for research on the statistical relationship of audio and video speech parameters with an audio-video (AV) automatic speech recognition (ASR) task in mind, but may be useful for other research tasks. AVOZES is the first published AV speaking-face data corpus for Australian English and is novel in its use of a stereo camera system for the video recordings and its modular design.

1. Introduction

The field of Audio-Video Speech Processing (AVSP) has received a growing interest by researchers around the world in recent years. With the advances in computer technology, ASR has become feasible in many application areas and environments, but still faces difficulties in the presence of acoustic noise. The use of the additional information of visible speech articulation can help to improve the recognition in noisy acoustic conditions. Over the past decade, various AV speech data corpora [1, 2, 3, 4, 5, 6] have been created to investigate methods for AV ASR, but many of these corpora lack a general design which makes the comparison of the results using different methods difficult.

A modular, extensible framework for the design of AV speech data corpora has recently been proposed by Goecke *et al.* [7]. In the design of a corpora, various factors need to be considered and addressed, such as typical speaker and task variables, environment and signal variables, and annotation and analysis variables. These factors are described in detail in Millar *et al.* [8]. The modular approach means that 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 access to the speakers, who appear 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 by explicit description. Mixing the content of modules creates an unwieldy design but well-designed modules that cover distinct areas facilitate a good corpus structure.

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

1. recording setup without a speaker,
2. recording setup with speaker,
3. coverage of phonemes and visemes.

The module "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 "recording setup with speaker" shows the speaker in the scene. This can include sequences useful or necessary for the video processing, such as views of the face from various angles. The module "coverage of phonemes and visemes" contains the basic speech components.

Additional modules can be easily added. Some modules, for each speaker, that were considered prior to the creation of the AVOZES data corpus described in this paper 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 specific acoustic or visible speech patterns (for example, lip rounding). These sequences can be used to classify speakers into different AV classes. 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 an idealistic data corpus design point of view, repeating the modules "recording setup with speaker" and "coverage of phonemes and visemes" for each different condition is desirable. Deviations from this ideal arise owing to constraints on resources and on the patience of speakers and should be noted in the corpus design description.

The remaining parts of this paper detail the design and recording setup of the AVOZES data corpus. Section 2 describes the modular design of AVOZES. Section 3 gives details on the recording setup. The recording process is described in Section 4. Section 5 contains the conclusions.

2. The Design of the AVOZES Data Corpus

The proposed framework was followed in the design of the *A*udio-*V*ideo *OZ*stralian *E*nglish *S*peech (AVOZES) data corpus [7]. No other AV speech data corpus with stereo camera video has been published thus far. Stereo vision can offer potentially more accurate measurements on the face than a single camera system can, as 3D coordinates can be recovered using the known stereo vision geometry of the recording system.

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

[†]Subsequent to this work, Goecke is now employed by National ICT Australia (NICTA). Millar is seconded to NICTA. NICTA is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

- the scene without any speaker;
- the scene with speaker, head turning;
- 'calibration sequences' exhibiting horizontal and vertical lip movements during speech production;
- CVC- and VCV-words in a carrier phrase covering the phonemes and visemes of Australian English;
- the digits "0"-"9" in a constant carrier phrase; and
- three sentences as examples of continuous speech.

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

2.1. Module 1 - Recording Setup without Speaker

This module contains one sequence in the AVOZES data corpus for each the two recording stages. It is a 30 second sequence of the recording scene viewed by the two cameras but without any speaker present. 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. Since the sequence in this module is speaker-independent, only one recording is needed. However, if corpus recordings were made over prolonged time spans (months or years), or in intervals, the sequence should be repeated once during each interval to record possible changes to the recording environment.

2.2. Module 2 - Recording Setup with Speaker

Head movements form an important part of many corpora, as it is rarely the case in practical applications that a speaker faces the camera in a full frontal view. The ability of tracking a moving face and measuring correct data of the visible speech articulation is therefore important. AVOZES contains a frontal view of the face as well as views on an angle of 45 to either side (5 seconds each). Module 2 contains one such sequence for each speaker.

2.3. 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 lipreading as well as AV ASR, the amount of visible speech articulation determines how much 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. 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 looking at 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 ..." and "e o e o e o ..." recorded in the AVOZES data corpus were each repeated continuously by each speaker for about 10 seconds. 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. Module 4 - Short Words in a Carrier Phrase to Cover Phonemes and Visemes

The sequences in this module form the core part of the AVOZES data corpus for statistical analyses of relationships between audio and video speech parameters. There are 44 phonemes (24 consonantal and 20 vocalic) and 11 visemes (7 consonantal and 4 vocalic) in AuE according to Woodward and Barber [9], Plant and Macrae [10], and Plant [11]. The phonemes can be categorised into 8 classes (Table 1). Similarly for the visemes, following [10, 11], there are 11 viseme classes (Table 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 1. Plant and Macrae [10] did not label their vowel and diphthong visemes but they are broadly:

- front non-open vowels and front close-onset diphthongs,
- open vowels and open-onset diphthongs,
- back / central non-open vowels and diphthongs containing these vocalic positions, and
- back / central open vowels and diphthongs containing these vocalic positions.

44 Phonemes in 8 groups		11 Viseme groups	
Oral stops	p, b, t, d, k, g	Bilabials	p, b, m
Fricatives	f, v, θ, ð, s z, ʃ, ʒ, h	Labio-dentals	f, v
Affricates	tʃ, dʒ	Inter-dentals	θ, ð
Nasals	m, n, ŋ	Labio-velar glides	w, r
Liquids and glides	l, r, w, j	Palatals	ʃ, tʃ, ʒ, dʒ
Short vowels	ɪ, ʊ, ɛ, ə, ɒ, ʌ, æ	Alveolar non-fricatives and plosives, velar plosives	l, n, j, h, g, k
Long vowels	iː, uː, əː, ɜː, ɔː, ɑː	Alveolar fricatives and plosives	z, s, d, t
Diph-thongs	eɪ, əʊ, ɔɪ, aɪ, aʊ, ɪə, ʊə,	Front non-open vowels and front close-onset diphthongs	iː, ɪ, ɛ, ɪə
		Open vowels, 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 1: *Phonemes and visemes of Australian English.*

The phonemes and visemes in the AVOZES data corpus were put in central position in CVC- or VCV-contexts 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/. The voiced bi-labial /b/ was used as the consonant context for CVC-words. The opening and closing of a bi-labial viseme clearly marks the beginning and end of the vocalic nucleus and thus facilitates the visual segmentation. Using /b/ instead of /p/ lengthens each word and thus gives more data to analyse. A bi-labial context causes strong coarticulation effects in the formants. However, these are quite predictable for /b/ and it is believed that the advantages of a bi-labial 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 bi-labial opening and closing before and after the word under investigation again helps with the visual segmentation process, in particular for the VCV-words. The prompts were presented together with pronunciation hints to the speakers during both 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 on how to pronounce the phoneme under investigation in that prompt.

2.5. Module 5 - Application Sequences - Digits

Digit recognition is a common task in automatic speech recognition and similar sequences can be found in a number of AV speech data corpora [2, 3]. The AVOZES data corpus includes for each speaker one sequence per digit, 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. 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 the digit recognition task.

2.6. 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:

- *"Joe took father's green shoe bench out."*
- *"Yesterday morning on my tour, I heard wolves here."*
- *"Thin hair of azure colour is pointless."*

Together with the first sentence, the second and third sentences were designed in such a way that they contain all phonemes and visemes of AuE.. 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 the continuous speech recognition task.

3. Recording Setup

The recordings took place in the audio laboratory of the Computer Sciences Laboratory (CSL) at the Australian National University and the same equipment and environment was used on both occasions. The CSL audio laboratory is a soundproof room with a small amount of background noise from the room's air-conditioning.

The speakers sat on a swivel chair in front of the stereo cameras, which were positioned with the help of a camera tripod. Using a swivel chair had the advantage of easy adjustment to different body heights and also facilitated the recording of the head turning sequences. 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, and removed 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). The distance from the face of a speaker to the cameras varied between 55-65cm. 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. The computer screen with the prompts was another 20cm (in horizontal direction) behind the cameras. Prompts were advanced per mouse click by the recording assistant, when the 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. The computer and DV recorder were housed in a padded cardboard box to reduce the amount of noise produced by the hard disk and cooling fans.

A clip-on microphone was attached to the speaker's clothes on the chest about 20cm below the mouth. The microphone was an omni-directional Sennheiser MKE 10-3 microphone. 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. The DV recorder was a JVC HR-DVS1U miniDV/S-VHS video recorder which also featured an IEEE-1394 DV in/out connector.

Figure 1: *Split-screen view of stereo facial image.*

The output of the stereo cameras was multiplexed into one video signal, then sent to a Hitachi IP5005 video card. The video signal was unscrambled, so that the video sequences on tape showed 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 (see Figure 1). 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 30Hz frame rate. Because of the way that the outputs from the two cameras were multiplexed, there was a 16.6ms delay between the output from each camera in any recorded video frame. In other words, it is possible that fast moving objects will be shown in two slightly different positions. While virtually undetectable by the human eye at normal video play rate (30Hz), 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).

4. Recording

Ten native speakers of Australian English were recorded over a period of one week in August 2000, and a further similar set of 10 speakers were recorded over a period of two days in August 2001. The same equipment, setup, and location were used on both occasions.

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. 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).

The group of speakers is gender balanced with 10 female and 10 male speakers. Six speakers wear glasses, two have beards. At the time of the recordings, these speakers were between 23 and 56 years old. The speakers were approximately classified into the three speech varieties of AuE 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. Still, the classification can be helpful for experiments and analyses that aim at identifying similarities and differences between the AuE speech varieties.

The AVOZES data corpus contains frontal face (10) recordings. The faces were illuminated from the front. Recordings were made for a clean audio condition. 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 recording assistant was present, so that speakers did not have to handle any of the equipment themselves and could concentrate on the speaking task. A total of 58 sequences were recorded per speaker (3 face model sequences and 55 speech material sequences).

5. Conclusions

AVOZES is the first AV speech data corpus published for Australian English. It follows a recently proposed framework for the design of such corpora that addresses the various factors relevant in AV corpus design [7, 8]. Another novel aspect is the recording with a calibrated stereo camera system, which potentially allows more accurate measurements by stereo reconstruction of 3D coordinates. AVOZES will be available through the Australian Speech Science and Technology Association (ASSTA, www.assta.org).

6. References

[1] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Boree, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication AVBPA2003*. Guildford, UK: Springer-Verlag, June 2003, pp. 625–638.

[2] C. Chibelushi, S. Gandon, J. 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*, London, UK, Digest Reference Number 1996/213, Nov. 1996, pp. 7/1–7/7.

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

[4] 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*, Washington (DC), USA, Mar. 1999, pp. 72–77.

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

[6] E. Patterson, S. Gurbuz, Z. Tufekci, and J. 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*, vol. 2. Orlando (FL), USA: IEEE, May 2002, pp. 2017–2020.

[7] R. Goecke, Q. Tran, J. 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 *Proc. 8th Australian Int. Conf. on Speech Science and Technology SST2000*, Canberra, Australia, Dec. 2000, pp. 92–97.

[8] J. 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.

[9] M. Woodward and C. Barber, "Phoneme Perception in Lipreading," *Journal of Speech Hearing Research*, vol. 3, no. 3, pp. 212–222, Sept. 1960.

[10] G. Plant and J. Macrae, "Visual Perception of Australian Consonants, Vowels and Diphthongs," *Australian Teacher of the Deaf*, vol. 18, pp. 46–50, July 1977.

[11] G. Plant, "Visual identification of Australian vowels and diphthongs," *Australian Journal of Audiology*, vol. 2, no. 2, pp. 83–91, 1980.