Aspects of Speaking-Face Data Corpus Design Methodology

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

[1]Australian National University, [2]University of Canberra, [3]Fraunhofer IGD-Rostock
Corresponding author: bruce.millar@anu.edu.au

Abstract

This paper develops a methodology for the design of audio-video data corpora of the speaking face. Existing corpora are surveyed and the principles of data specification, data description and statistical representation are analysed both from an application-driven and from a scientifically motivated perspective. Furthermore, the possibility of "opportunistic" design of speaking-face data corpora is considered.

1. Introduction

The design of corpora for audio-video (AV) speech studies and technology development has not yet been evaluated to the same extent as the design for audio-only corpora. The design methodology greatly affects the usefulness of the corpora depending on both the rigour and generality with which it is implemented. As more AV data corpora are collected, it is appropriate for researchers and developers to ask how they may be re-used. The mapping of the needs of one project onto the data corpus used for an earlier project is complex and will be the more successful the better the design of that corpus is specified.

A methodology can be developed from the perspective of future applications such as AV automatic speech recognition (ASR), AV person authentication and speaking-face synthesis, which may be embedded in such systems as cash dispensers, videoconferencing, 4G mobile phones, door access control etc. Such applications will require thorough scientific experimentation to determine their likely performance and their optimal designs. This, in turn, requires data corpora that are designed to answer the scientific questions in a sound statistical manner, such as by providing a representative sample of the target population for the proposed applications.

A second perspective for the corpus design methodology is based on descriptions of the physical and behavioural features that characterise the speaking-face. The audio and video features relating to the speaker and the speaking environment together with limitations caused by instructions to the speaker and by audio and video detection and measurement constraints comprise the main body of relevant descriptive data. It is important, for example, to know that features are represented at a resolution to enable relevant distinctions to be made.

A third perspective for corpus design is the question of targeted design versus opportunistic design. A targeted design methodology aims at facilitating a particular set of experiments to answer a specific question or support a specific application – or perhaps a family of applications. Opportunistic design, on the other hand, aims at the utilisation of existing data, such as television broadcasts, and aims to facilitate useful scientific experimentation by careful selection of already available data. Targeted design is invariably more expensive, but opportunistic design may not provide all the required scientific answers. The two standpoints are not mutually exclusive. They can be linked by equivalent rigour in data description since good targeted design will anticipate future opportunistic use of a corpus, while good opportunistic design can facilitate specific experiments on the basis of already available data.

In this paper we address AV corpus design from these three stances, examine the design of published corpora and discuss methodological issues for future AV corpus design.

2. Data Specification

The obvious starting point for the design of an audio-video data corpus of the speaking face is to establish the purpose of the data collection. This may be done by the careful design of a scientific experiment or by the analysis of the anticipated applications whose development will depend on those data or which will be facilitated or enhanced by the availability of such a data corpus.

The utility of an audio-video database, as distinct from a single-modality corpus, lies in the information contained in each of the audio and video modalities as well as in the interaction between the acoustic speech data and the corresponding facial configurations. Therefore, the experimental design will define the variables of interest in both modalities much as would be the case for single-modality data collection, and, in addition, will address issues such as the synchronisation between the audio and video signals, joint probabilities between the single-modality variables and the statistical coverage of the combined audio-video parameter space.

When an audio-video data corpus of the speaking face is collected in support of specific applications, the design of the corpus depends critically on the type of application, such as

- speech recognition applications enhanced by the simultaneous transmission of the facial image of the speaker,
- multimodal biometric person authentication, for forensic or non-forensic purposes,
- speech synthesis applications enhanced by the simultaneous transmission of a synthetic facial image,
- audio-video coding applications for the low-bit rate transmission of the speaking face;
- and a range of possible application aiding the communication of both the hearing and vision-impaired.

Depending on the specific experimental design or envisaged application, the design of the audio-video data corpus will entail detailed data definitions in the four categories of speaker, speaking task, speaking environment, and signal measurement/processing. Finally, the utility of the corpus will depend substantially on the annotation and analysis provided for the collected data.

3. Data Description

A structured analysis of audio [1] and AV [2] descriptors of complex speaking scenes has been presented by Millar. Here

we summarise the entities, environments, measurements and processes that generate useful descriptors of the total AV scene.

The speaker is the essential entity in the speaking scene. The speaking face is broadly characterised by its static appearance, the dynamic activity of the vocal organs and the interaction of the two. These may be represented by parameters describing the habitual settings of facial and vocal organs that determine the speaker's range of image and sound possibilities. The configuration of chin, lips, nose and eyes provides the visual cues of the speaking activity, whereas the existence of facial hair, spectacles and aspects of skin colour limit access to these cues. Long-term evidence of vocal tract length and internal muscle tone guide the interpretation of acoustic output (see Table 1).

The speaking task is defined in terms of the language, linguistic content and context of the speech, as governed by the speakers' instructions and prompts (see Table 1).

FACTORS	SOME DESCRIPTORS		
Appearance	Facial configuration	Facial hair	Spectacles, skin colour(s)
Vocal organs	Tract length	Vocal tract settings	
Habit	Hesitation	Head & eye movement	
Task	Language	Content	Prompt

Table 1. Typical Speaker and Task Variables

The speaking environment includes the position(s) and orientation(s) of the speaker(s) and of all audio and video transducers. The acoustic and visual characteristics of the environment will comprise the size and reverberation of the space plus any background acoustic sources, and illumination sources plus visual background (see Table 2).

The signal measurement descriptors include the type and settings of all transducers, and for each transducer, any channel transformations, such as colour balancing or audio filtering, between the transducer and eventual data storage. Digitisation parameters and the format in which this raw data is stored must also be included. In some cases the digitised signal will be compressed according to a protocol that needs to be defined and its parameters given (see Table 2).

FACTORS	SOME DESCRIPTORS		
Topology	Positions	Orientations	
Ambience	Ac. noise	Illumination	Background
Transducer	Type	Settings	
Transforms	Process	Parameters of process	

Table 2. Typical Environment and Signal Variables

Annotation & Analysis processes are critical for the processing of raw audio-video speech data. The human processes of manual annotation should be accompanied by clear descriptions of the human expertise, the theoretical basis for judgements, the quality of the representation of the data provided for the human annotator [3], and the expertise of transcription checkers. Automated analysis can be described by the identification of generic analytic processes and their parameters, but also the description of heuristic constraints are very important in such areas as formant tracking (see Table 3).

FACTORS	SOME DESCRIPTORS		
Expertise	Training	Experience	Practice
Theory	Reference to specific theory		
Quality	Raw data	Display	Checking
Analysis	Process	Parameters	Heuristics

Table 3. Typical Annotation & Analysis Variables

4. Statistical Validity

A significant design issue is the adequacy of the data corpus to allow valid generalisations of the data. Users may need a representative and balanced coverage of the phonemes, visemes and allophones of the target language for speech recognition or a representative and balanced coverage of the speech and facial characteristics of the target population for authentication. A representative data corpus in this sense may be obtained by collecting a large number of samples from the target population and by ensuring that the samples are random. On the other hand, valid representation may also be achieved by a smaller number of samples and careful analysis of the variations in the population data, which may, for example, take account of the well-known "sheep-and-goat" problem in statistical speech and speaker recognition, i.e. the fact that some people show variability in their speech that is poorly represented in the majority of speakers.

5. Opportunistic Data Collection

"Opportunistic" data collection refers to the recording of data that is available to be recorded, but is not necessarily open to modification according to any design strategy on the part of the data collector. A typical example is the recording of free-to-air news broadcasts, which is capable of providing a large volume of AV speech data for a range of scientific experiments and potential applications.

While opportunistic corpus design does not offer the same degree of data definition as a targeted corpus, it can be designed by way of a set of decisions that determine which of the recordings and parts of recordings are retained in the corpus and described by maximising the use of the descriptors outlined in Section 3. Such data collection, if well designed and adapted for the scientific or developmental purpose at hand, can be an extremely time and cost-efficient means of collecting required data.

6. Specific Application-Related Issues

There are a number of issues in the design of a speaking-face data corpus that are related to particular applications or families of applications. All of these fall into one or another of the categories of speaker, speaking task, environment and signal, which were discussed at length above. We do not attempt to provide an exhaustive list of such issues here, but illustrate the kind of consideration that is often required by way of some examples:

Example 1: A data corpus for the development of an AV speech recognition system requires representative samples of all the speech sounds of the target language(s) including the dialects and foreign accents present in the target population and a representative sample of male and female subjects with a random distribution of variables such as age and educational level. Depending on what kind of signal detection is planned for the application, the corpus may require a representative set of samples from a range of webcams and PC microphones for internet-based applications, or a representative set of samples from a range of mobile phone sets, connection conditions, environmental noise and illumination conditions for a fourth-generation mobile-phone based application.

Example 2: A corpus in the domain of face-voice authentication requires that sufficient data be captured for each individual "client" so that a reliable statistical model can be constructed for the average characteristics of that individual's speech sounds and facial appearance and for the likely variability of those characteristics. In the speech domain such data would need to include a representative set of phonemes, words and phrases that would be used for authentication, while in the facial domain the data would need to include the likely variations of the face, such as hair style, beard and eye glasses. Further data may be required to allow the modelling of such variables as environmental noise, telephone line characteristics, rhino-laryngeal conditions, facial size, tilt and azimuth, and incident lighting, as well as the longitudinal change of both voice and facial characteristics, measured over time spans that are appropriate for the application.

Example 3: The volume of daily television broadcast news lends itself to be "designed" into an opportunistic data corpus of the speaking face – either for speech recognition or for face-voice authentication. In Canberra, there are currently 5 free-to-air TV stations with 25 daily news broadcasts, a local cable television provider carries a further 7 international channels with hourly news programs around the clock, and a 3-metre satellite dish delivers more than 30 additional stations with regular news programmes. For authentication research, a speaking-face corpus can be recorded from these news programmes with between 4 and 10 newsreaders per station per week and between 5 and 50 sessions per newsreader per week. A limited amount of the spoken content, namely "the station announcement" is relatively constant between the sessions of one station, but the majority of the content is variable. Data specification is limited to the selection of the material available and data description is generally limited to the long-term AV analysis of the newsreader and a post-hoc analysis of the visual and acoustic ambience of the studio and the signal mix.

7. Design of Existing AV Corpora

In a survey of AV speech corpora, Chibelushi *et al.* [4] examined existing corpora as well as the features that researchers would like to see in such corpora. Based on the responses from a questionnaire that was sent out to other researchers, existing AV speech corpora were typically found having only a small number of speakers, covering only a small number of phonemes and visemes, and containing isolated words (digits, letters of the alphabet) rather than embedded or continuous speech. 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' case,
- whole-face images in colour,
- short words and continuous speech with transcription, and
- extensibility.

While more AV speech corpora have been created since this survey, many of the considerations still apply and agree with the issues raised in this paper. The remainder of this section compares AV speaking-face corpora of reasonable size (≥ 20 subjects, utterances covering a variety of factors named in Section 3; see the summary in Table 4).

The BANCA database [5] was specifically recorded for research in authentication. Recordings were made both with a webcam (low quality) and a digital camera (high quality), as well as low and high quality microphones, which addresses speaking environment and signal measurement factors. Each recording contains utterances of a random 12-digit number, the subject's name, address, and date of birth as samples of the speaking task variable 'content'. Recordings were made in controlled, degraded, and adverse scenarios (total of 12 sessions) but seem to include only different visual backgrounds, not different audio conditions.

Other corpora for authentication purposes are the M2VTS [6] and XM2VTS [7] databases, which also serve as databases for lips-speech synchronisation and AV ASR research. Recordings contain utterances of the digits 0-9, one sentence of continuous speech, and head rotation sequences, with each subject being recorded in several sessions. These sequences are therefore very useful for determining the speaker variables.

The DAVID BT database [8] was recorded with both authentication and AV ASR tasks in mind. It consists of 4 subcorpora (SC). SC-1 contains sequences for face segmentation tasks with variable visual scenarios (illumination, facial distracters, background), in which the subjects utter the digits 0-9, thus sampling the speaker variable 'appearance' and the environment variable 'ambience'. SC-2 was recorded for AV ASR and authentication tasks. The speakers utter the digits 0-9. A subset of the speakers has highlighted lips (blue make-up). SC-3 contains VCVCV utterances for research in speech-assisted video compression and synthesis of talking heads. SC 4 contains sentences from a business control set for AV ASR and identity recognition for video-conferencing. Both frontal and profile views (using one camera and a mirror) were recorded. These last three SC give good samples of the 'task' variable.

The proprietary IBM AV LVCSR corpus [9] contains continuously spoken utterances from the IBM ViaVoice training set, recorded in two scenarios (car, office) using different equipment (low and high quality cameras and microphones) with partly different speakers (one session only). The main purpose is large vocabulary continuous AV ASR research. This corpus samples the variables 'task', 'appearance', 'ambience', and 'transducer' particularly well.

The CUAVE corpus [10] contains recordings of isolated and connected digits (phone number spelling task) for AV ASR research, with the subjects either standing still or moving (sampling the variable 'topology' particularly well). An extra level of confusion is added by pairs of speakers being recorded in some sequences. Sequences are fully labelled at a millisecond level.

The VidTIMIT database [11] was designed for AV person verification and contains 43 speakers, each recorded over 3 sessions speaking phrases from the NTIMIT corpus [12]. Appearance, head rotation and camera zoom factor were varied.

The AVOZES corpus [13] was designed using a targeted approach to enable scientific experimentation on the AV characteristics of Australian English speech. AVOZES contains six modules. Modules 2-6 were recorded for each speaker:

1. the scene without any speaker;
2. the scene with speaker, and with head rotation;

Corpus	Language	Subjects	Video	Audio kHz/bits	Sess-ions	Applications	Environment
AVOZES	En	10f, 10m	NTSC DV	48 / 16	1	AV ASR	Studio conditions
BANCA	En/Fr/Sp/It	4×(26f, 26m)	PAL DV	32 / 16,12	12	Authentication	Diff backgrounds
CUAVE	En	17f, 19m + 20 pairs	NTSC DV, MPEG-2	44 / 16	1	AV ASR	Subjects moving, pairs of subjects
DAVID SC1 SC2 SC3 SC4	En	2f, 5m 61f, 62m 2f, 3m 61f, 62m			1 5 1 5	Face segmentation AV ASR, authentication Compression, synthesis AV ASR, authentication	Variable visual scenarios, some speakers with highlighted lips
IBM AV	En	290	MPEG-2	22 / 16	1	AV ASR	Car, office
M2VTS	Fr	37	Hi8, CIF	48 / 16	5	Authentication, AV ASR	Const conditions
VidTIMIT	En	19f, 24m	PAL DV, JPEG	32 / 16	3	AV ASR, authentication	Noisy office
XM2VTS	Fr	295	PAL DV	32 / 16	4	Authentication, AV ASR	Const conditions

Table 4. Overview of Existing Audio-Video Speaking-Face Data Corpora

3. `calibration sequences' exhibiting extent of horizontal and vertical lip movements during speech production;
4. CVC- and VCV-words in a carrier phrase covering the phonemes and visemes of Australian English;
5. the digits "0"-"9" in a constant carrier phrase; and
6. three sentences as examples of continuous speech.

AVOZES is novel in being recorded with a calibrated stereo camera system, thus offering 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 modular design allowed extensibility with respect to some of the data description issues raised in Section 3 of this paper. Module 1 is a sample of the speaking environment variable 'ambience'. Module 2 samples the speaker variable 'appearance'. The sequences in module 3 can be useful for determining the speaker variable 'habit' as well as some analysis variables. Module 4 contains the core sequences of the corpus, which sample the phonetic space and would be of most interest for ASR research. Modules 5 and 6 are application-driven utterances (e.g. number-spelling task) and potentially of interest for ASR and authentication research.

8. Conclusions

The paper has emphasised the need for a methodical design of audio-video data corpora of the speaking face. Whether the motivation of corpus collection is to support scientific research or specific applications, whether the design is targeted or opportunistic, the utility of the corpus is affected to a large degree by the rigour of data specification and data description.

9. References

[1] J.B. Millar, "A structure for comprehensive spoken language description", in *Proc. First Int. Conf. on Language Resources and Evaluation (ICLRE'98)*, Granada, Spain, May 1998, Vol.2, pp. 1303-1308.

[2] J.B. Millar, "Customisation and quality assessment of spoken language description", in *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP'98)*, Sydney, Australia, Nov 1998, Vol.4, pp. 1575-1578.

[3] J.B. Millar, "Labelling the Labellers", COCOSDA 1996 Workshop, Philadelphia, USA, and at http://rsise.anu.edu.au/csl/hci/labelling_the_labeller .

[4] C.C. Chibelushi, F. Deravi, and J.S. Mason, "Survey of audio visual speech databases", Tech. Rep., Department of Electrical and Electronic Engineering, University of Wales, Swansea, UK, 1996.

[5] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, J.-P. Thiran, "The BANCA Database and Evaluation Protocol", in *Proc. 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication AVBPA2003,* Guildford, UK, 2003, Springer-Verlag.

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

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

[8] 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*, London, UK, Digest Reference Number 1996/213, Nov. 1996, pp. 7/1-7/7.

[9] 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.

[10] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy, "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research", in *Proc. ICASSP2002*, Vol. 2, pp. 2017-2020.

[11] C. Sanderson and K.K. Paliwal, "Fast Features for Face Authentication under Illumination Direction Changes", in Pattern Recognition Letters, Vol. 24, No 14, 2003, pp. 2409-2419.

[12] C. Jankowski, A. Kalyanswami, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced Continuous-Speech Telephone Bandwidth Speech Database", Proc. ICASSP-1990, Vol. 1, pp. 109-112.

[13] R. Goecke, J.B. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES", submitted to Interspeech 2004 – ICSLP, Jeju, Korea, Oct. 2004.