CAVA (human Communication: an Audio Visual Archive)

Project Report

File formats for use in the CAVA repository

Matt Mahon
CAVA Project Officer
July 2009

http://www.ucl.ac.uk/ls/cava

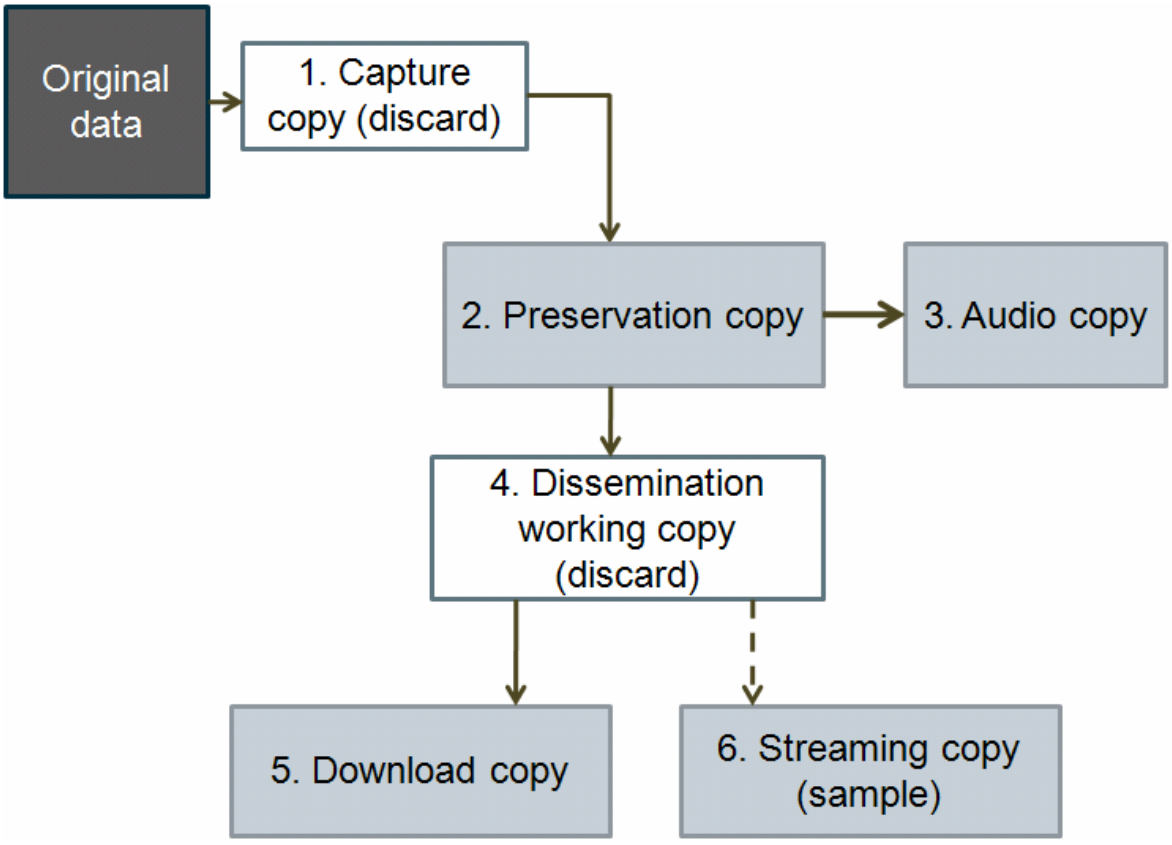
1. Introduction

The CAVA project will build a repository to collect and share audio-visual datasets created by UCL's Human Communication researchers. This short document summarises the file formats to be used in the CAVA repository.

2. CAVA data

The data which will be deposited in the CAVA repository comes from a wide range of sources and in a wide range of formats, with an equally wide range of software requirements. The project aims to introduce uniformity of format wherever practicable. Each recording will be held in three versions: a preservation master format, not publicly available, and two access formats, compressed video and audio-only. Additionally, small sample videos will be made available in streaming format at 'collection' level, to help potential users to explore the repository and identify the datasets most appropriate to their work.

3. CAVA file creation workflow: overview



4. File formats

4.1. Preservation format

For CAVA purposes, a container file like an AVI contains four objects: video data, audio data and video and audio codecs. The codecs provide the information about how the data are stored within the file and how to uncompress them for viewing. A program which is able to correctly identify and open a container file might not be able to subsequently decode the actual data stored within, because either the metadata in the container file are not sufficient, or the software lacks the specific codec, to interpret the actual data the file contains. So a container format is only as useful as the codec it is packaged with, and the technical aspects of the video and audio data it contains can be altered as appropriate. The capture files from Mini-DV have a data rate of 28.8mbps, which is very large (streaming videos, for instance, tend to have a data rate of around 300kbps). These files are unwieldy, but will rarely be accessed. Preservation files from Mini-DV, which make up the majority, will maintain their original data rate. The DV codecs used in the original files will often be proprietary codecs which are built into the hardware used to record the videos. All DV-AVIs will be converted to an industry standard DV25 codec. At this stage it is appropriate to apply a filter to remove deinterlacing artefacts.

Originals from other formats may have differing data rates, and every effort will be taken to maintain their quality. Low-quality legacy data, such as those which only exist in MPEG-1 format, will not be preserved at higher resolutions or data rates. The aim is to simply maintain the data. The audio will remain uncompressed. By these means the full quality of the original, whatever that may be, will be preserved.

4.2. Dissemination: video

The CAVA video download format must be one which is small enough to download and manipulate, and which has a widely-recognised codec. MPEG-1 video with an MPEG-1 layer II audio codec (MP2) is the most appropriate, given that the dissemination copies must work on managed computers. A data rate of around 3024kbps is preferred, in order to strike a balance between high-quality preservation copies and manipulable downloads. Where lower quality MPEG-1s exist, as with some legacy data, a minimum data rate of 1000kbps is acceptable. The dissemination files are suitable for analysis using analytical software such as DHCS Videolab, CLAN (CHILDES), Transana, Elan and Praat, among others.

4.3. Dissemination: audio-only

The audio-only files to be produced by CAVA are uncompressed (PCM), in WAV format, 44100Hz, 16-bit stereo with a data rate of 1024kbps, which is industry standard CD quality. Given the relatively small data rate of high-quality audio, there is no reason not to maintain the highest quality possible here. Note that audio data from natural conversation may occasionally be unsuitable for speech

processing software, where it was not recorded using appropriate hardware or formats.

4.4. Dissemination: streaming format

The UCL video streaming server uses files in Flash Video format (FLV), with an ON2 VP6 video codec. The data rate is 400kbps. This is an aggressive compression, but one which does not significantly reduce the subjective quality of the data. Both the download video and the streaming video use lossy compressed audio. As MP2 and MP3 are standards for the file types, there is little that can be done to change this. However, as the sample videos are intended to give the user a taster of the content, rather than the quality, of the datasets, this is acceptable.

5. Summary of CAVA format specifications

		Capture	Preservation	Download	Streaming	Audio-only
	File type	AVI	AVI	MPG	FLV	WAV
Video	Codec	[DVSD]	DV25	MPEG-1	On2 VP6	N/A
	Data rate (kbps)	28800	28800	3024	400	N/A
	Frames/sec	25	25	25	25	N/A
	Frame size	720x576	720x576	720x576	480x360	N/A
Audio	Codec	PCM	PCM	MP2	MP3	PCM
	Data rate (kbps)	1024	1024	224	128	1024
	Sampling rate (Hz)	44100	44100	44100	44100	44100
	Channels	2	2	2	2	2
	Sample precision	16-bit	16-bit	16-bit	16-bit	16-bit