# CAVE: AN ON-LINE PROCEDURE FOR CREATING AND RUNNING AUDITORY-VISUAL SPEECH PERCEPTION EXPERIMENTS - HARDWARE, SOFTWARE, AND ADVANTAGES

*Denis Burnham, John Fowler, and Michelle Nicol*

**School of Psychology, University of NSW, Sydney, 2052, Australia.**

**Tel. +61 2 9385 30 25, Fax: +61 2 9385 36 41, email: d.burnham@unsw.edu.au**

## ABSTRACT

The McGurk effect or fusion illusion, in which mismatched auditory and visual speech sound components are perceived as an emergent phone, is extensively used in auditory-visual speech perception research. The usual method of running experiments involves time-consuming preparation of dubbed videotapes. This paper describes an alternative, the Computerised Auditory-Visual Experiment (CAVE), in which audio dubbing occurs on-line. Its advantages include reduced preparation time, greater flexibility, and on-line collection of response type and latency data.

## 1. INTRODUCTION

There has been a recent resurgence in the study of humans' auditory-visual speech perception [1,2]. Much of this research employs the McGurk effect, in which for example, auditory [b] dubbed onto visual [g], is perceived by human adults and children as [d] or [Δ] (as in 'them') [3]. This effect is pervasive, occurring for various consonants [4,5] in various vowel contexts [6], and for vowels [7]. As the McGurk effect occurs in good viewing and listening conditions, it demonstrates that humans automatically use visual information for speech perception whenever it is available. Beyond mere demonstration, it has been used as a research tool to show the phonetic nature of auditory-visual integration [6, 8], for studying infants' auditory-visual speech perception [4,5,9,10], and to investigate cross-linguistic factors in auditory-visual speech perception [11,12].

In these McGurk effect studies participants are presented with various stimulus combinations: auditory-only (AO) (motionless face with dubbed sound), visual-only (VO) (appropriate lip movements without sound), matching auditory-visual presentations (AV-m), and mismatching auditory-visual (AV-mm) presentations - either McGurk presentations, eg, auditory [b] and visual [g], or combination presentations, eg, auditory [g] and visual [b]. Most studies of the McGurk effect have involved presenting pre-recorded dubbed videotapes to participants whose task it is to provide spoken or written responses. This has a number of shortcomings: videotape dubbing of auditory components onto visual components is time-consuming, and exact auditory-visual synchronisation may be difficult to achieve [though see 13]; data cannot usually be collected on-line; reaction times are difficult to measure [though see 8]; and each experiment requires preparation of new dubbed videotapes. The CAVE overcomes these problems.

## 2. THE CAVE

The essence of the Computerised Auditory-Visual Experiment (CAVE) is that the analog signal from the audio channel of a videotape acts, through a voice key, as a digital trigger for the output of a pre-programmed auditory component digitally stored on disk. The audio output on the videotape is not heard, and the onset of the audio output from disk occurs as near as possible to *exactly* the onset of the original sound on the videotape. Thus, as shown in Table 1, for AV trials a sound from disk replaces the original sound. For consistency, the AV-mm and AV-m auditory-visual trials are created by the same procedure: on AV-mm trials an auditory component *different* to the visual component is programmed; while on AV-m trials the *same* auditory component as the visual component is programmed. On VO trials an audio file containing 'silence' is programmed to play with the visual component. For AO trials a tone-marker positioned on the second audio channel triggers a sound from disk. As the procedure is computer-controlled, data can be collected on-line through digital response key inputs, and as a clock is started when the audio input from the videotape signals that a sound from disk is to be played, reaction times (RTs) can be collected. In addition, the procedure

| | **Video** | **Audio** Channel 1 | **Audio** Channel 2 |
|---|---|---|---|
| *Auditory-Visual Trials* | Auditory-Visual Recording | Voice triggers sound from disk | -- |
| *Visual-Only Trials* | Auditory-Visual Recording | Voice input triggers 'silence' | -- |
| *Auditory-Only Trials* | Still Face; No Voice | -- | Tone triggers sound from disk |

***Table 1: Video and Audio Components in Auditory-Visual, Visual-Only, and Auditory-Only Trials***

can be used to run on-line experiments in which be created for use in off-line experiments, eg, when the computer apparatus is not available, when it is necessary to run a number of subjects simultaneously such as in a quick pilot study, or when the same stimulus material is to be used in different laboratories. Moreover, the program used to create such experiments and videotapes is generative, so new experiments can be created easily, and modified through a parameters file.

| | | | |
|---|---|---|---|
| | **MAKEFUS generates either a reaction time (FUSION) or a no reaction time (NoRTFUS) experiment called *filename*<br>These are implemented by:** | | **MAKETAPE filename** |
| | ↓ **FUSION filename** | ↓ **NoRTFUS filename** | Creates dubbed videotape from either FUSION or NoRTFUS Experiment |
| *Features* | On-line dubbing<br>Fixed responses<br>Response key input<br>Response type & latency data<br>On-line data collection<br>FUSREP → data collection<br>Participants tested individually | On-line dubbing<br>Open set responses<br>Written or oral responses<br>Response type data<br>Enter data after testing<br>FUSREP → data collection<br>Participants tested individually | Participants tested individually or in groups.<br>Data may be entered later using FUSREP |
| *Advantages* | On-line collection of both response type and latency.<br>Automatic data collation | No pressure on participants for speeded responses | Permanent and portable videotape for testing in different labs |

***Table 2 Software Components, and Features and Advantages***

### 2.1 Software

To run an experiment a stimulus person is videotaped producing all the required auditory-visual stimuli. The auditory components from the audio channel of the videotape are then digitally recorded, and stored on disk. The software components are set out in Table 2, along with the features and relative advantages of three implementations - FUSION, NoRTFUS, and MAKETAPE. A generative interactive program, MAKEFUS, is used to set out the characteristics of the experiment, the nature of each trial, the names of sound files, etc. The method of data collection is also chosen - either a response panel with up to six appropriately-labelled response keys can be used for digital input, or participants simply write down or enunciate their open set responses. Completion of the MAKEFUS interaction results in the production of an executable program, *filename.exe*, and an editable parameters file, *fuspara.txt*, with associated sounds, raw data, and collated data sub-directories.

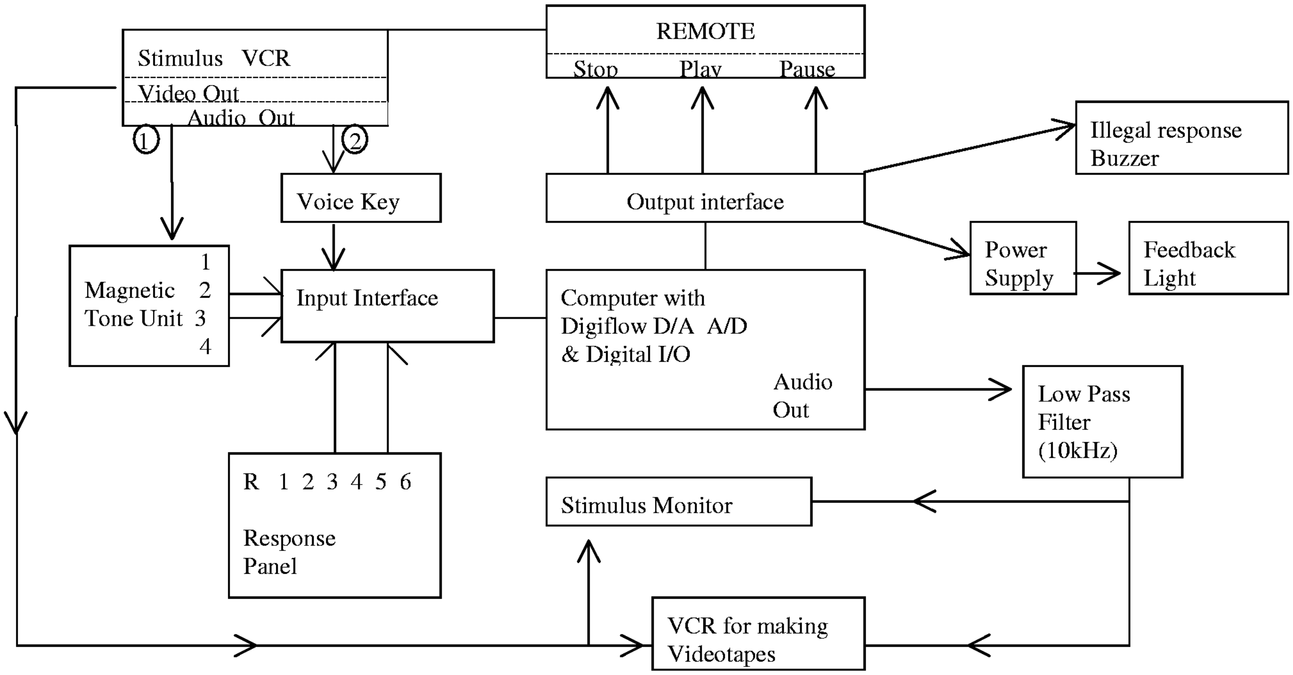
The response panel reaction time option is run by FUSION *filename.* In such an experiment, response type and latency are recorded by the subject depressing one of the six response keys. For example in an auditory [b] visual [g] experiment these response keys may be labelled "b", "g", "d", "th", "bg", and "gb". Trial control is provided by means of a seventh key on the response panel labelled “ready”. If this is not depressed within a specified period after a response key choice then the videotape pauses until it is. The No RT option is run by NoRTFUS *filename*. Such an experiment is run on-line but participants make written or oral responses. Finally, either of these types of experiments can be used to create videotapes using MAKETAPE *filename*. The output of this operation is a videotape or videotapes, with the dubbings specified in the file created by MAKEFUS. Such videotapes are more quickly produced than by videotape dubbing, and incorporate the accuracy inherent in the CAVE. Moreover, fresh copies of the videotapes can be made without loss of quality inherent in adding an extra videotape generation.

Data collation is handled by FUSREP, which collates participants data and adds it to the data already collected. In FUSION experiments raw data is stored on-line ready for collation; in NoRTFUS experiments data can be input interactively using a FUSREP option after testing each participant or at the end of the experiment.

### 2.2. Hardware

The hardware configuration is shown in Figure 1. The *computer* is a 386 or better, and is equipped with an A/D - D/A and digital I/O board (*Digiflow* in this case). The *remote control* unit of the stimulus VCR is doctored so that digital inputs are able to control the ‘Play’, ‘Stop’, and 'Pause' functions. The *filter unit* houses a 10 kHz low pass filter to eradicate high frequency noise. The *voice key* is a custom-made unit with adjustable upper and lower amplitude thresholds to allow all and only all speech tokens on Channel 1 of the stimulus VCR to activate digital input. The magnetic tape *tone unit* can be used both to generate and respond to tones placed on Channel 2 of the stimulus VCR. Four different tones (each of a different frequency) can be used. In the current set-up these different tones are used to signal to the executable the (i) start of a tape, (ii) start of a block, (iii) start of an AO trial on which no sound component was originally recorded, and (iv) end of a tape.

***Figure 1: Hardware Configuration***



### 2.3. Experimental Parameters

Details of software manipulations are set out below.

*Subject Characteristics:* In both FUSION and NoRTFUS experiments subject details - name, age, and sex are stored along with the data.

*Groups:* Often different groups of subjects are run in an experiment. These may be differentiated on the basis of subject variables, such as language background or age, which require no change in the FUSION or NoRTFUS procedures; or between-group manipulations such as the stimulus speaker, or sound file characteristics, which *do* require changes in FUSION or NoRTFUS procedures. Such variations are automatically implemented by the selection of the subject group at the onset of the experiment. Data from up to 16 groups can be included, and are collated separately by FUSREP.

*Conditions:* Up to three different conditions can be included in an experiment. These are within-group manipulations, and may include variations of major parameters, eg, sound file names to allow cross-language or cross-speaker face/voice matches and mismatches, or variations of the vowel context for consonant decisions. Up to six phones (auditory and/or visual components) can be included in each condition.

*Practice Phase:* Practice trials may be included either just at the beginning of the experiment, or at the beginning of each condition. Up to four blocks of 24 trials may be included. In FUSION experiments we typically use the practice phase to present stimuli which will elicit responses on all the designated response keys. Thus the practice phase facilitates participants' full use of the response possibilities, as well as their familiarity with the response procedure. Inclusion of a significant proportion of AO and especially VO trials in this phase ensures that participants learn to watch the video monitor on all trials. The practice phase also aids in establishing fast and accurate responding: if a response occurs after a specified period after sound onset, the *illegal response buzzer* sounds and the experimenter can use this to remind subjects that quick RTs are required. The *feedback light* option for correct responses is also typically used in the practice but not the test phase. The results for the practice phase are stored along with each participant's data but separate from their test trial data.

*Test Phase:* Up to four blocks of 48 trials may be included in a test phase. Typically we use this phase to present the AV-mm trials of interest in the experiment, along with various AO, VO, and AV-m trials.

*Blocks:* Blocks in the same condition are identical in constitution (numbers of each trial type). However, the order of trials between blocks may vary. The start of a block (and a condition) is determined by the programmed parameters, but a check can be provided by inserting a start of block tone marker (Tone 2) on the videotape. Titles can be inserted on the videotapes between blocks (and conditions) to allow subjects to rest, or to provide them with encouraging feedback.

*Trials:* On each trial a phone is presented in one of four manners - AO, VO, AV-m, or AV-mm. Any trial duration and any inter-trial interval (ITI) is possible and permissible, and both may vary across trials, because for trial initiation the executable program simply relies on a digital input triggered by the voice on Channel 1 of the stimulus VCR, or by a Tone 3 input on AO trials. In addition, after selection of a response key on a particular trial, the videotape is paused and onset of the next trial is delayed until the participant depresses the ready key. The sequence of events on the original videotape determines the base trial duration and ITI. We typically employ 4 sec trial sequences, constituted as follows: 2 secs of motionless face, a 1-sec envelope in which the phone is presented, 1 sec of motionless face, and 1 sec of blank tape.

*Exemplars:* Between one and five exemplars of each audio phone may be included on disk. Exemplars may be induced so that a random selection of one of the sound files of the format *soundA?.bin*, where ? is a number between 1 and 5, is played whenever *soundA* is specified in the parameters. This option encourages participants to disregard the acoustic variability between exemplars of the same phone and respond just to the phonetic invariance of the sound. (Ideally this should be used with a system by which the visual exemplars of the phone are also varied.) In another exemplar option all trials in a particular block have the same exemplar. This may be used, for example, where speaker identity varies between blocks. Finally, exemplars may be separately specified for each trial to include manipulations such as cross-pairing of a particular speaker's audio and video exemplars, or to mix auditory and visual exemplars.

## 3. CONCLUSIONS

The CAVE, incorporating a hardware configuration (Figure 1) and a suite of programs, MAKEFUS, FUSION, NoRTFUS, MAKETAPE, and FUSREP (Table 2) provides a number of advantages over the usual method of running fusion experiments. These include:

- Relatively quick and efficient preparation of new experiments or pilot tests
- Reduced videotape preparation time (no dubbing required)
- Accurate (on-line) dubbing of visual and auditory components either for on- or off-line experiments
- Storage of experimental parameters for future replications of experiments
- Storage of experimental parameters for construction of identical videotapes without loss of videotape quality by an added generation
- Real-time collection of data including reaction times
- Data collation in a form ready for statistical analysis

## 4. REFERENCES

[1] D.G. Stork & M.E. Hennecke (Eds.) *Speechreading by humans and machines.* Springer-Verlag: Berlin.

[2] R. Campbell, B. Dodd & D. Burnham (Eds) *Hearing by Eye Part 2: The psychology of speechreading and auditory-visual speech.* Psychology Press: Hove, England.

[3] H. McGurk & J. McDonald (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

[4] D. Burnham (1992) Processing auditory-visual speech in infancy & across phonologies. *Int. J. Psych*, *27*, 59.

[5] D. Burnham & B. Dodd (1996) Auditory-visual speech perception as a direct process: The McGurk effect in infants & across languages. In D. Stork & M. Hennecke (Eds.)

[6] K. Green (1996) The use of auditory and visual information in phonetic perception. In [1] D.G. Stork & M.E. Hennecke (Eds.)

[7] J. Robert-Ribes, M. Piquemal, J-L Schwartz & P. Escudier (1995) Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In [1] D.G. Stork & M.E. Hennecke (Eds.)

[8] K.P. Green & P.K Kuhl (1991) Integral processing of visual place and auditory voicing information during phonetic perception. *J. of Exp.Psych: HP&P*, *17*, 278-288.

[9] J.A. Johnson, L.D. Rosenblum & M.A. Schmuckler (1995) The McGurk effect in infants. *J. Acoust. Soc.Amer. 97*, 2aSC7. 3286.

[10] R.N. Desjardins & J.F. Werker (1996) 4-month-old female infants are influenced by visible speech. Poster presented at the Inter. Conf. of Infant Studies, Providence RI.

[11] K. Sekiyama & Y. Tohkura (1991) McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *J. Acoust. Soc.Amer*, *90*, 1797-1805.

[12] D. Burnham (1997) Language specificity in the development of auditory-visual speech perception. In [2] R. Campbell, B. Dodd & D. Burnham (Eds.)

[13] M. McGrath & Q. Summerfield (1984) Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *J. Acoust. Soc.Amer*, 77, 678-685.

## ACKNOWLEDGMENTS

The second author was responsible for hardware design and, the third for software design and implementation. Peter Wyatt wrote real-time software, and Peter Boorman and Gordon Muir constructed hardware components. Megan Smith, Elizabeth Francis and Sheila Keane assisted in testing the CAVE, and Vee Scott assisted with manuscript preparation.