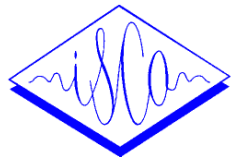

Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

Why captions have to be on time

Burnham, D.[(1)]*, Robert-Ribes, J.*[(2)] *and Ellison, R.*[(3)]

[(1)] School of Psychology, University of NSW, <d.burnham@unsw.edu.au>
[(2)] CSIRO-Mathematical and Information Sciences, < Jordi.Robert-Ribes@cmis.csiro.au>
[(3)] Australian Caption Centre, <caption@wr.com.au>

ABSTRACT

Closed captioning dramatically improves deaf people's enjoyment of television shows, and appears to augment the auditory signal for people with some degree of hearing impairment. However, reports from people with mild to severe hearing loss suggest that when there is a delay between the audio track and the caption, perceivers are confused unless they turn down the volume. These effects have not yet been investigated experimentally. This study provides a preliminary investigation of the importance of synchronisation of captions with auditory-visual material for hearing-impaired people's enjoyment and comprehension of captioned television programs. Two participants were presented with audio-caption delays of 0, 1, 2, and 4 secs in an auditory-visual condition and an auditory-only condition. Both enjoyment and intelligibility diminished over lag times. In general enjoyment and intelligibility were higher in the auditory-visual than the auditory-only condition, however, for the more severely hearing impaired of the two participants, both enjoyment and intelligibility diminished at a faster rate over delay times for the auditory-visual than the auditory-only condition. Thus at long delays the presence of the visual signal appeared to be distracting. These results are discussed in terms of perceptual mechanisms and practical applications for captioning.

1. INTRODUCTION

Almost 30 years ago the first demonstrations of closed captioning took place in the United States. Since then, closed captioning has generated a great deal of enthusiasm in deaf and hearing impaired people. The difference between closed and open captions is a technical one. Closed captions are broadcast as part of the program and the viewer needs a caption decoder to see them. On the contrary, open captions are always on the screen; there is no need of special equipment to see them. They look similar to subtitles on foreign language films. From the point of view of the perception of the captions, closed and open captions are identical. For this study we used closed captions, yet it can be assumed that the same results would have resulted from using open captions.

What is captioned television? In captioned TV programs the soundtrack of the program is shown as text on the TV screen. Captions may be coloured and positioned on the screen to show each character's speech. Sound effects, music and other audio cues are incorporated to ensure all relevant information is available to the viewer. This is the main difference between captions and subtitles, because subtitles simply give a translation of the dialogue into another language [6]. Figure 1 shows an example of a TV caption.

Figure 1: Example of caption

Who uses captioned television? Deaf people rely solely on captions to follow television programs or commercials when lipreading is not possible. Hearing impaired people with some degree of hearing use captions to help understand the soundtrack. However, not only people with hearing impairment use captions. Captions are also widely used to watch television programs in noisy environments (sporting clubs, pubs, etc) and research interest has been generated for using captions for teaching handicapped students or foreign language students ([2], [7]).

Anecdotal reports from television viewers with mild to severe hearing loss suggest that captions facilitate the perception of auditory material. This evidence is mainly negative in nature: for instance, if words are *incorrectly captioned*, hearing impaired perceivers notice the discrepancy; and when there is a *delay* between the audio track and the caption perceivers report that they sometimes turn the volume down. However, neither of these claims has been systematically studied in controlled experiments.

Since captions were first introduced, captioning techniques have improved immensely. Nevertheless, there are still cases in which the desynchronisation of the audio-visual material and the captions occurs. One good example of such cases is live captioning when a stenographer types the captions on-line, as is usually done for TV news programs. In this case there is the normal delay involved in the typing so that the captions appear some time after the soundtrack. An example of such a delayed caption is provided in Figure 2. As suggested above hearing impaired perceivers may find such delayed captioning annoying or confusing. In this preliminary study the effect of audio track-caption delay on intelligibility and subjective pleasantness was investigated systematically with two hearing impaired perceivers.

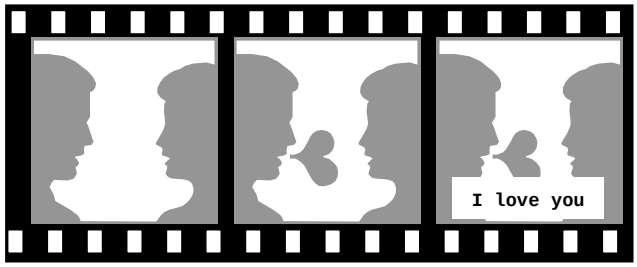

Figure 2: Example of a delayed caption

2. METHOD

There were two hearing impaired participants in the study. Both were males and both were 71 years of age at the date of testing. The first, DM, had severe hearing impairment contracted early in life (maybe at birth) but not diagnosed until he was 16 years old. He has no auditory ability with high frequency sounds but some limited ability with low frequencies. He reported that he watched captioned TV for an average of 2 hours per day, and always left the sound on when he watched captioned TV. The second participant, RR, had a more recently contracted moderate high frequency loss, and had been wearing a hearing aid just for the last 10 years. He reported that he watched television about 3 to 4 hours per day.

The stimulus materials consisted of two prerecorded videotapes, the Audio-Video (AV) tape and the Auditory-Only (A-only) tape. For the AV tape, captions accompanied a normal videotape with auditory and visual information; whereas on the A-only tape only the audio track and not the visual track were present. This was included to ascertain whether the unpleasantness associated with delayed caption conditions occurred due to the mismatch of voice and captions (A-only condition) or face/voice and the captions. Each tape contained 12 captioned news segments, of 15 to 35 secs duration. The audio track-caption delay (A-C Delay) on these segments varied: of the 12 excerpts on each tape, there were 2 excerpts with 0 secs delay, i.e., perfect captioning; 4 with 1 second A-C Delay; 4 with 2 seconds A-C Delay; and 2 with 4 secs A-C Delay. Thus a Modality (AV vs A-only) x Delay (0, 1, 2, 4 seconds) design was employed in the study. Both participants completed all conditions.

The study was conducted in a quiet room at the Australian Caption Centre [1]. Both participants were tested at the same time. The AV tape was presented first, followed by the A-only tape. The participants were required to complete three tasks on each trial: a rating of the excerpt for 'pleasantness'; a rating for 'intelligibility'; and a brief written description of the content of the excerpt. For the first two, the participants were asked to provide a rating from 0 to 100, where 0 indicates extremely bad or totally unintelligible captioning, and 100 extremely good or perfectly intelligible captioning. The written description was required mainly to ensure that participants concentrated on the content of each excerpt. Before testing, participants filled out a brief questionnaire, and the experimenter explained the testing procedure. In addition, an initial trial was presented from a TV program known to have not very pleasing captioning. This was given in order to provide the participants with practice at using the scale, and in order to provide a baseline for the later ratings.

3. RESULTS

Each participant's ratings of pleasantness and intelligibility were averaged across A-C Delays to produce four lag scores (0, 1, 2, and 4 secs) for each of the two tape conditions (AV and A-only). These are graphically presented in Figure 3 for each of the two participants for each of the two measures, pleasantness and intelligibility. As can be seen, pleasantness and intelligibility diminished over lag times, although the degree and rate of decrease appeared to be greater for pleasantness ratings. In general, ratings on both measures were higher in the auditory-visual than the auditory-only condition. For RR, the participant with the better hearing, the AV and the A-only conditions were more similar than they were for DM. In addition for RR there is no drop-off in intelligibility or pleasantness until *after* 1 second of delay.

On the other hand for DM, the participant with the greater hearing impairment, the drop-off occurs as soon as any delay is introduced, and the relative pleasantness of the AV and A-only conditions depends upon the delay. Pleasantness ratings

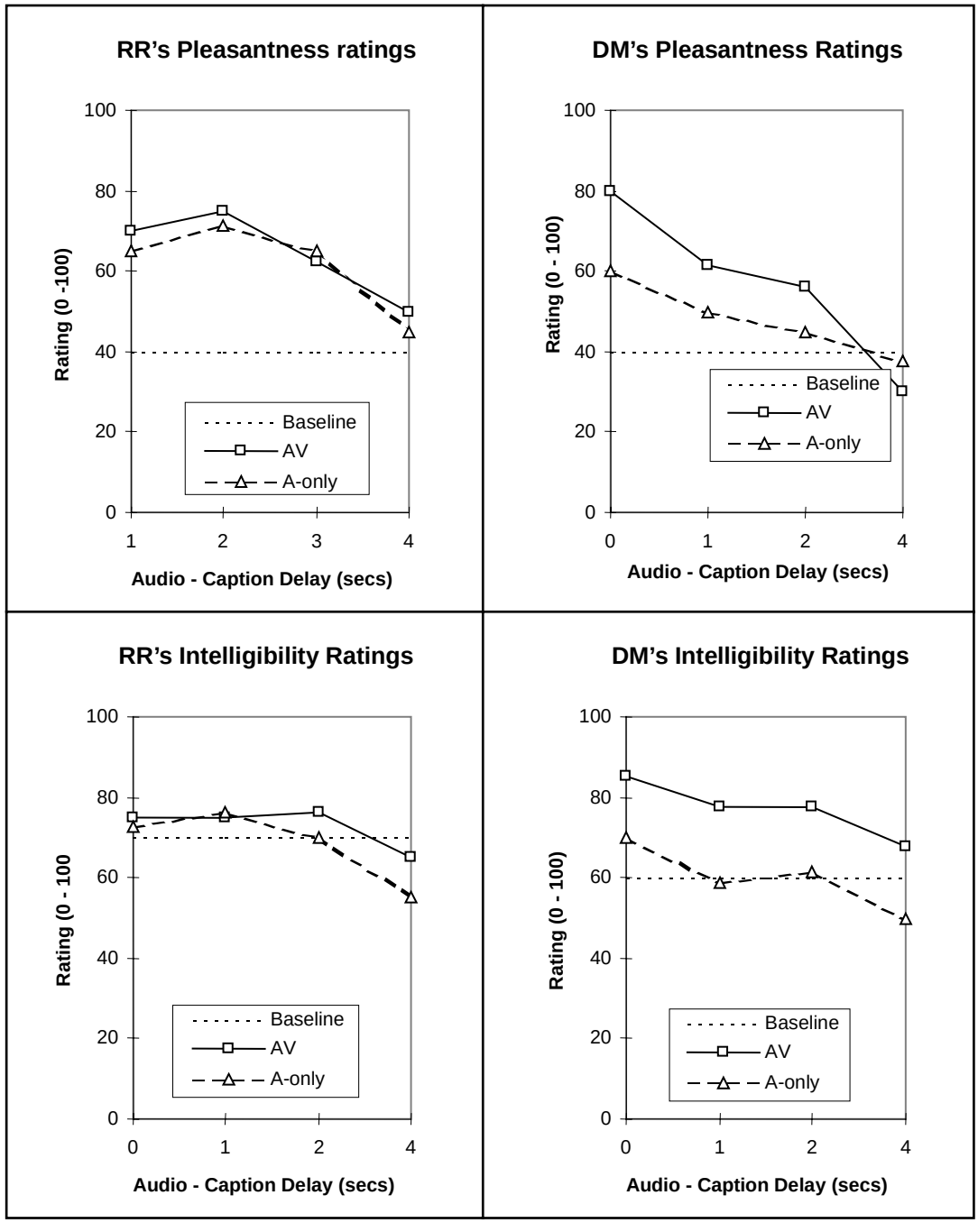

Figure 3: Rating of pleasantness and intelligibility for each subject

diminish at a faster rate over delays for the AV than the A-only condition, such that at long delays there is a crossover. AV is judged to be more pleasant than A-only at 0, 1, and 2 secs delay, but less pleasant at 4 secs delay. Thus the presence of the visual signal with long lags appears to be distracting. Intelligibility, however, for DM is not similarly affected (the decrease for AV and A-only parallel each other), so while at long delays the presence of the visual signal appears to be unpleasant or distracting, it still aids intelligibility.

4. DISCUSSION

The results appear to show that the integration of linguistic information from the auditory system and from written language is disrupted by the asynchrony between these channels. A similar effect has also been shown for the asynchrony between auditory and visual signals (lip and face movement) and found that there was a disruption of integration once the delay exceeded 80 msecs (for a

review see [3] and for a discussion see [5]). This disruption has a ceiling effect at 280 msecs ([4]).

As disruption for RR only began sometime after 1 second in our study, it may be the case that the human system is more sensitive to delays of information when the input is from a common source, i.e., speech articulations (as in [3] and [4]), than when it is from two basically discrete sources: articulated speech, and written language. However, the test employed by [3] and [4] may be more sensitive than that used here, so in future studies it may be useful to devise a more sensitive test of audio caption delay to complement perceivers' ratings. In such studies it would also be useful to determine whether misperceptions or illusions occur when audio and written speech do not match, either temporally or phonologically.

In the data presented here there is preliminary evidence that the more severe the hearing impairment, the greater is the effect of audio-caption delay. If this effect is reliable, and this is yet to be determined in further studies, then there could be various reasons for it. On the one hand, one might expect that the better a person's hearing, the more they would be distracted by the asynchrony. On the other hand, it may be the case that for such people with less severe hearing loss, captions are only used as a secondary source to check information if auditory input isn't entirely clear to them. People with a greater degree of hearing impairment might then rely more on the caption, while still processing the speech to whatever level is possible for them. In order to investigate this more effectively it would be of interest in future studies to test the effect of the complete absence of the auditory signal. This could be done in one of two ways. The first is to test completely deaf participants. On the basis of the data presented here it would be expected that such subjects would be very sensitive to audio - caption delays. Another means to a similar end would be to introduce a condition in which there is no auditory signal, such that the delay would simply be between the visual speech and the captions. To the extent that the participants rely upon visual speech, this should be distracting for them. Relative ratings of pleasantness and intelligibility under such conditions would be of interest, in the light of DM's differential decrease for pleasantness and for intelligibility in the AV condition here.

It appears clearly from our results that captioning companies must ensure good alignment of the captions to the sound track if they want the viewers to keep the interest in watching the captioned programs. Even delays of one second seem to disrupt the enjoyment and comprehension of a television program. There are, of course, practical problems with the production of perfectly synchronous speech and captions under all conditions, and these must also be considered.

This preliminary study of delayed captioning raises interesting questions at both the theoretical and practical levels. These should be followed up in future studies.

5. REFERENCES

1. Australian Caption Centre, http://www.auscap.com.au/
2. Thorn, F. and Thorn S. (1996). "Television captions for hearing-impaired people: A study of key factors that affect reading performance". Human Factors, 38(3), 452-463.
3. Summerfield, Q. (1992). "Lipreading and audio-visual speech perception". In V. Bruce et al. (Eds.), Processing the facial image (71-78), Oxford, Clarendon Press.
4. Smeele P. and Sitting A. (1991). "The contribution of vision to speech perception". Proc. 2nd European Conf. On Speech Communication Technology (1495-1497), ESCA, Genova (Italy).
5. Abry, C., Cathiard, M.A., Robert-Ribes, J. and Schwartz, J.L. (1994). "The coherence of speech in audio-visual integration", Current Psychology of Cognition, 13(1), 52-59.
6. De Linde, Z. (1995). "'Read my lips', subtitling principles, practices and problems". Perspectives: Studies in translatology, 3(1), 9-20.
7. Koskinen P.S., Knable, J.E., Markham, P.L., Jensema, C.J. and Kane, K.W. (1996). "Captioned television and the voaculary acquisition of adult second language correlational facility residents". J. Educational Technology Systems, 24(4), 359-373.

6. ACKNOWLEDGEMENTS

The authors wish to acknowledge Martin Curnow and Sheila Keane for their help in preparing the video material.