6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HEARING BY EYE: VISUAL SPATIAL DEGRADATION AND THE MCGURK EFFECT

Dr John MacDonald, Soren Andersen, Dr Talis Bachmann
Department of Psychology, University of Portsmouth, Portsmouth, Hampshire, PO1 2DY, UK
(e-mail: john.macdonald@port.ac.uk; talis.bachmann@port.ac.uk)

ABSTRACT

McGurk and MacDonald (*Nature* **264**, 746-748, 1976) discovered that when a discrepancy is created between visual information from lip movements and speech information from the auditory channel then perceivers often report a percept that is neither the auditory or visual stimulus, an illusory response. This 'McGurk' effect is strong evidence that perceivers extract key information about a speech sound from concomitant visual articulation. This study investigates the effects of spatial quantisation on the McGurk effect. Participants (N=20) were presented with incongruous auditory-visual combinations of simple consonant vowel tokens. The visual stimulus was intact or had undergone various degrees of degradation through spatial quantisation. McGurk type responses were significantly influenced by levels of quantisation with more veridical auditory responses at the coarser levels of quantisation. However, even at the coarsest level of quantisation some McGurk type responses were reported.

10.21437/Eurospeech.1999-301

INTRODUCTION

The issue of how different modalities interact with one another to produce a unified and integrated perceptual environment is most clearly seen in the area of audio-visual speech perception [4, 11]. A number of studies have demonstrated greater intelligibility of speech when lip movement information is available [18]. However, in addition to facilitation there is also evidence for more integrated processing occurring. The most convincing demonstration of this is the McGurk effect [14, 10]. When participants are presented with conflicting information from each of the two modalities, e.g. visual /ga/ with auditory /ba/ the perception of what is heard is either an illusory fusion, i.e. a heard /da/ or a combination, e.g. /bga/ where the lip movements are for /ba/ and the auditory stimulus is /ga/. This is a robust finding which has been confirmed by a number of studies [16, 8, 12, 21, 15].

Although this interactive influence of visible lip movements on auditory speech perception is well established, it is not clear what mechanisms and processes govern the resultant perception. Research has established that the most informative areas of the face are those surrounding the lips, including the jaw and cheekbone [17, 3]. Also eye movement studies have shown that although most fixations of gaze are to the eyes and mouth, the proportion of the duration of the gaze being fixated on the mouth increases as masking noise levels increase [20].

A technique that has recently been used to study what information is detected from faces and how it is used to make perceptual judgements is that of spatial quantisation [1, 7, 5, 2, 19]. To date, only one study has used this technique to investigate visible speech perception [6]. They found that speech reading performance was relatively resistant to the effects of image degradation, although the effect varied across different speech tokens. The rationale behind this technique is to systematically introduce different levels of spatial degradation so that the levels of local feature analysis will be relatively more impaired than the levels of more global configuration analysis, and it should be possible to find out the relative impact of these different levels on the perceptual process under investigation.

The present study seeks to apply this approach to the study of audio-visual speech perception using the McGurk effect to auditory-visual conflicting stimuli as the key measure. It is hypothesised that as the McGurk effect depends on the processing and influence of the visual stimulus, and as spatial degradation reduces the information derived from the visual component then the relative frequency of illusory responses should decrease with increasing levels of degradation.

There are two main aims: (1) to measure the effects of quantisation in order to find out the *tolerance* of the McGurk effect to degradation of the visual stimulus and (2) to answer the question of whether there is a gradual decline of the influence of the facial information with coarseness of quantisation or whether there is a critical point at which information is lost resulting in a *quantum decrease* of the illusion. Previous research with face identification has found that with the gradual increase in coarseness of quantisation, i.e. increase in the size of the pixels, there is a dramatic drop of face recognition at a certain critical value of quantisation as measured by the number of pixels/face [1, 5].

METHOD

Participants

20 native speakers of British English, 10 males and 10 females, aged between 18 and 40, with normal or corrected to normal vision and normal hearing took part in a 2 hours session.

Stimuli

The face of a young woman was videotaped uttering pairs of consonant-vowel (CV) syllables. Eight pairs of CV syllables were used: /ba-ba/, /da-da/, /ga-ga/, /pa-pa/, /ta-ta/, /ka-ka/, /ma-ma/, and /na-na/. Each CV syllable pair was spoken three times with an interval of approximately 1 second between repetitions, producing thus a triple of pairs, e.g. /ba-ba/, /ba-ba/, /ba-ba/. To ensure visibility of the oral cavity and to eliminate shadows in the face, spotlights were placed at an angle of 45°, 2 metres to the left and right of the speaker. For each recording the speaker's face was positioned at the centre of the camera. She was instructed to articulate the syllables naturally, to avoid artificial emphasis, to close her mouth between repetitions, and to sit as still as possible. Three recordings of each CV syllable triple resulted in a total of 24 master stimuli. From these master stimuli, congruent and incongruent AV speech examples were created from within the following consonants groups; [b d g] (voiced), [p t k] (voiceless) and [m n] (nasal). This procedure gave a total of 22 AV speech combinations; 14 incongruent and 8 congruent. The stimuli were produced by dubbing the selected auditory stimulus onto the selected visual stimulus. To ensure coincidence of the auditory and the visual signal during the release of the consonant in each utterance, single frame editing was used.

Degrading the stimuli

The spatial quantisation (area averaging mosaic tansformation [9]) of the AV stimuli was performed by using a Panasonic WJ-MX30 mixing/special-effects board and two Panasonic AG7350 video-recorders. From pilot observations and previous speaking face recognition research with quantised stimuli [6], five levels of spatially quantised AV stimuli were prepared: 0 (the original high-resolution image at the TV monitor level of resolution), level 3 (29.2 pixels/face, 11.6 pixels/mouth), level 5 (19.4 p/face, 7.7 p/mouth), level 7 (14.2 p/face, 5.6 p/mouth) and level 9 (11.2 p/face, 4.4 p/mouth). The five levels of spatial quantisation were crossed with the 22 AV speech stimuli, yielding a total of 110 spatially quantised AV stimuli.

Blocking the stimuli

Each trial consisted of 5 seconds of the face prior to articulation and 1 second of the face after articulation, with a 10 second inter-trial interval. Two pseudo-random sequences of the 110 trials were created, avoiding sequential repetitions of AV stimuli.

Procedure

The AV stimuli were presented on a 20" monitor, from a S-VHS Panasonic-AG7350 video recorder. Participants sat on a chair approximately 100-110 cm from the screen, fixating the central part of the screen at eye level. The vertical size of the face presented on the screen, measured at spatial degradation level 0, was approximately 28 cm. The horizontal width of the face, from cheekbone to cheekbone was approximately 18.0 cm. The horizontal width of the mouth when the lips were closed was approximately 7.0 cm. At a viewing distance of 100-110 cm, the face subtended 16 vertical and 10.3 horizontal degrees of visual angle, with the mouth subtending 4.0 horizontal degrees of visual angle. All testing was conducted individually in a sound attenuated room. Participants were instructed that they were to watch a video of a woman model uttering some meaningless, but intelligible syllables, and that their task was to report what they heard the model say. Each participant experienced each of the two different sets of sequences of 110 stimuli, i.e. 220 trials in total. Participants received £10 as payment.

RESULTS

The main dependent variable was the number of correct auditory identification responses which was calculated for each participant for each quantisation condition and each incongruent audio-visual stimulus combination. The overall results are presented in Table 1. The higher the value the weaker the McGurk effect demonstrating audio-visual interaction. The maximum score was 2.0 which represents completely veridical auditory responding on both trials. As can be seen from Table 1 the rate of correct auditory responding was influenced by the auditory-visual stimulus combination ($F(9,171) = 10.12$, $p<0.001$) and by level of spatial quantisation ($F(4,76) =24.74$, $p<0.001$). There was also a significant interaction between stimulus pair and quantisation ($F(36,684) = 5.54$, $p<0.001$). In order to explore these effects a series of subsidiary analyses of variance were carried out. For example it is clear that for many speakers and for listeners to those speakers the phonemes /d/ and /g/ are visually indistinguishable. There should be little difference between responses to the pairings of auditory /ba/ - visual /da/ and the pairing of auditory /ba/ - visual /ga/. This was tested directly by comparing the number of correct responses across these conditions and across levels of quantisation.

Visual - /ba/; auditory - /da/ or /ga/

There was no significant effect of the auditory stimulus and no significant interaction with level of spatial quantisation. However, there was a significant main effect of spatial quantisation ($F(4,76) = 2.90$, $p<0.05$). Participants made fewer errors at coarser levels of quantisation.

Visual - /pa/; auditory - /ta/ or /ka/

There was no main effect of the auditory stimulus used but there was a significant main effect of quantisation ($F(4,76) = 3.22$, $p<0.05$) and a significant interaction between auditory stimulus and quantisation ($F(4,76) =4.29$, $p<0.005$). This later interaction resulting from

Table 1 Mean number of correct auditory responses as a function of visual context, auditory stimulus and spatial quantisation level (Standard Deviations in parentheses, N=20).

Visual	/ba/	/ba/	/da/	/ga/	/pa/	/pa/	/ta/	/ka/	/ma/	/na/
Auditory	/da/	/ga/	/ba/	/ba/	/ta/	/ka/	/pa/	/pa/	/na/	/ma/
Quantisation Level										
0	1.80	1.70	1.05**	1.00**	1.90	1.55*	1.10**	1.00**	.95**	.25**
	(.52)	(.66)	(.99)	(.97)	(.31)	(.76)	(.97)	(.97)	(.83)	(.64)
3	1.85	1.70	1.20**	1.25**	1.95	1.80	1.10**	1.20**	.75**	.80**
	(.49)	(.66)	(.95)	(.91)	(.22)	(.62)	(.97)	(.95)	(.72)	(.95)
5	1.90	1.65*	1.15**	1.35**	1.90	1.95	1.40**	1.45**	1.20**	1.25**
	(.31)	(.75)	(.93)	(.88)	(.31)	(.22)	(.88)	(.89)	(.83)	(.91)
7	1.95	1.95	1.55*	1.50**	2.00	2.00	1.80	1.75	1.65*	1.75
	(.22)	(.22)	(.83)	(.76)	(.00)	(.00)	(.52)	(.44)	(.49)	(.64)
9	1.95	1.85	1.70	1.40**	1.90	1.95	1.85	1.50**	1.55*	1.80
	(.22)	(.49)	(.66)	(.88)	(.31)	(.22)	(.37)	(.69)	(.60)	(.52)

* - (Mean = 2.0) $p<0.05$, ** - (Mean = 2.0) $p<0.01$

more errors at the finer levels of quantisation for the auditory /ka/ stimulus.

Visual - /da/ or /ga/; auditory - /ba/

There was a significant main effect for quantisation level ($F(4,76) = 8.87$; $p<0.001$) but no other significant effects. The effect of quantisation was that participants made fewer errors at the coarser levels of quantisation.

Visual - /ta/ or /ka/, auditory - /pa/

The only significant effect was for quantisation level ($F(4,76) = 25.64$; $p<0.001$) which as before was fewer errors at the coarser levels of quantisation.

Visual - /ma/, auditory - /na/ versus visual - /na/, auditory - /ma/

Here again there was no effect of stimulus pair, but only a significant main effect of quantisation level ($F(4,76) = 31.76$; $p<0.001$) and significant interaction between Stimulus Pair and Quantisation ($F(4,76) = 4.49$; $p<0.01$). The interaction being more errors for visual - /na/, auditory - /ma/ than the reverse pairing for the undegraded stimulus, with no differences between the pairs for the degraded stimuli.

In conclusion it can be seen that previous results regarding the level of audio-visual interaction are supported by the data here. For the stop consonants there is more interaction when bilabial sounds - /ba/ and /pa/ are combined with velar and alveolar lip movements - /da/, /ga/, /ta/ and /ka/ than when they are combined in the opposite order. Also there is no difference in visual impact of certain phonemes, e.g. /da/ and /ga/, and /ta/ and /ka/.

In order to assess the impact of the levels of quantisation on reported auditory perception a series of single mean 't' tests were carried out comparing the mean number of correct auditory responses with the maximum possible correct that could be achieved, i.e. a value of 2. This was done for each level of quantisation and for each audio-visual stimulus pair. It was previously established through analysis of auditory only presentation and congruent audio-visual stimulus combinations, e.g. auditory - /pa/ with visual - /pa/, that the auditory stimuli were perceived accurately. There were 2400 auditory-visual congruent or auditory only trials. Only 4 of these trials had error responses and these were random across participants and stimuli therefore the auditory stimuli were perceptible when presented either on their own or in a congruent audio-visual context. Table 1 shows which values differed significantly from 2. The results show a clear effect of quantisation across all stimulus combinations. There are fewer error responses at the coarser levels of quantisation and this holds across all stimulus pairings. However, the level of audio-visual interaction varies across stimulus combinations with more pronounced interaction for nasal phonemes and for bilabial auditory stimuli dubbed onto non-labial visemes.

CONCLUSIONS

The main results of the study are clear. The McGurk effect [15, 10] was replicated. The incongruous visual speech information resulted in a significant number of errors in auditory perception. These effects were particularly pronounced when bilabial auditory stimuli (/ba/, /pa/ and /ma/) were dubbed onto velar or alveolar visual stimuli (/da/, /ga/, /ka/, /ta/ or /na/). The typical 'fusion' type error responses to these stimuli were alveolar (/da/, /ta/) or nasal (/na/) syllables. In contrast to previous research there were relatively few combination type responses (/bga/, /kapka/ or /mna/) to velar and alveolar auditory stimuli dubbed onto bilabial visual stimuli.

In addition there were strongly significant effects of quantisation, especially for the incongruent stimuli that resulted in 'fusion' type responses. Some combinations only showed significant 'McGurk' effects at levels 0, 3 and 5 whereas the stimulus pairs visual /ga/ - auditory /ba/ and visual /ka/ - auditory /pa/ showed significant effects at level 9. At level 9 only very coarse information regarding the movement of the lips and jaw can be discerned in the visual stimulus. Therefore in response to our first question the answer is that the McGurk effect is very tolerant of degradation of the visual stimulus, at least for some specific stimulus pairings. Regarding the issue of whether there would be some quantum decrease in reporting the illusion or a gradual decline, again the answer is quite clear cut. There is no evidence in this data to support the proposal of a critical quantisation point at which information is suddenly lost. In contrast, the data support the idea of a gradual decline in error responses, perhaps as a result of a loss of information along some continuum.

How then do these data fit with current theories of speech perception and what do they bring to the debate between competing theories? Some accounts emphasise the importance of articulatory features as the common metric for phonetic information being provided from the auditory and visual information [13]. However, these theories rely on the extraction of categorical segments which are then combined. It is not clear how this kind of theory would incorporate the gradual decline in visual influence evident in the data presented here. In contrast, Massaro's FLMP model [11] is based on continuously varying information being extracted from the auditory and visual modalities which is then mapped onto underlying phonetic representations. This approach may well be able to account for the gradual decrease in visual influence found here. However, further research is needed to investigate the differences found across the different pairings of syllable and modality. Further work is also needed on whether other response judgements can be used in addition to verbal identification. Understanding these differences may then possibly begin to show whether this integration of auditory and visual information is unique to speech or is part of a broader set of processing mechanisms and our basic perceptual capacity.

REFERENCES

1. Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology, 3,* 87-103.

2. Bachmann, T., & Kahusk, N. (1997). The effects of coarseness of quantisation, exposure duration, and selective spatial attention on the perception of spatially quantised ("blocked") visual images. *Perception, 26,* 1181-1196.

3. Benoit, C., Guiard-Marigny, T., Le Goff, B. & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In D.G. Stork & M.E. Hennecke (Eds), *Speechreading by Humans and Machines: Models, Systems and Applications* (pp. 315-28). NATO ASI Series, Berlin: Springer.

4. Bertelson, P. (1998). Starting from the ventriloquist: The perception of multimodal events. In M. Sabourin, F.I.M. Craik, & M. Robert (Eds.), *Advances in psychological science, II: Biological and cognitive aspects*. (pp. 419-439). Hove: Psychology Press.

5. Bhatia, S.K., Lakshminarayanan, V., Samal, A. Welland, G.V. (1995). Human face perception in degraded images. *Journal of Visual Communication and Image Representation, 6,* 280-295.

6. Campbell, C.S., & Massaro, D.W. (1997). Perception of visible speech: Influence of spatial quantization. *Perception, 26,* 627-644.

7. Costen, N.P., Parker, D.M., & Craw, I. (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Perception and Psychophysics, 58,* 602-612.

8. Easton, R.D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception and Psychophysics, 32,*562-570.

9. Harmon, L.D., & Julesz, B. (1973). Masking in visual recognition: Effects of two-dimensional filtered noise. *Science, 180,* 1194-1197.

10. MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics, 24,* 253-257.

11. Massaro, D.W. (1998). *Perceiving talking faces. From speech perception to a behavioral principle.* Cambridge, Mass.: Bradford/MIT Press.

12. Massaro, D.W., & Cohen, M.M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 753-771.

13. Mattingly, I.G. & Studdert-Kennedy, M. (Eds). (1991). *Modularity and the motor theory of speech perception*. Hillsdale, NJ: Erlbaum.

14. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

15. Munhall, K.G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception and Psychophysics, 58,* 351-362.

16. Roberts, M., & Summerfield, A.Q. (1981). Audio-visual adaptation in speech perception. *Perception and Psychophysics, 30,* 309-314.

17. Summerfield, Q. (1979) Use of visual information for phonetic perception. *Phonetics*. 36, 314-331.

18. Summerfield, A.Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London, B, 335,* 71-78.

19. Uttal, W.R., Baruch, T., & Allen, L. (1997). A parametric study of face recognition when image degradations are combined. *Spatial Vision, 11,* 179-204.

20. Vatikiotis-Bateson, E., Eigsti, I.M., Yano, S., & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics, 60,* 926-940.

21. Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception and Psychophysics, 57,*1124-1133.