INVESTIGATING THE ROLE OF LUMINANCE BOUNDARIES IN VISUAL AND AUDIOVISUAL SPEECH RECOGNITION USING LINE DRAWN FACES

M.V. McCotter[1] *& T.R. Jordan*

School of Psychology; University of Nottingham
University Park, Nottingham, U.K

ABSTRACT

Two experiments are reported which investigate the contribution of luminance boundaries to visual and audiovisual speech perception using colour, grey scale and line drawn talking faces. Use of line drawn faces should isolate basic luminance boundaries, while removing the distribution of luminance from the face. Unimodal auditory and visual syllables were combined to produce congruent (matching) and incongruent (McGurk) speech stimuli. Visual speech presented in line drawn faces was highly recognisable and influenced perception of congruent and incongruent auditory speech. However, visual speech presented in colour and grey scale faces was slightly more accurate and influential on perception of auditory speech perception than visual speech presented in line drawn faces. In light of these findings, the role of luminance boundaries in perception of visual and audiovisual speech is discussed.

INTRODUCTION

Although speech recognition is often regarded as primarily an auditory process, auditory speech can be influenced strongly by the appearance of a talker's face [1]. A powerful demonstration of the influence of visual speech is shown by McGurk effect [2] whereby different auditory and visual inputs combine to form an illusory percept (e.g., when auditory /ba/ is presented with the facial movements /ga/, observers often report hearing 'da' or 'tha').

It is commonly (and logically) assumed that basic visual cues must be encoded for visual speech to be recognised and to affect auditory speech recognition in face-to-face communications [3], [4], [5]. However, recent findings suggest that detailed information (e.g., skin pigmentation, skin texture,) is not critical for visual and audiovisual speech recognition [6], [7], [8]. For example, comparisons between colour and grey scale talking faces [6] suggest that luminance information (present in colour and grey scale talking faces) may be a particularly powerful component in visual and audiovisual speech perception. Indeed, preliminary findings suggest that when the distribution of luminance across the face is disrupted by presenting talking faces in photo-negative, perception of visual speech is less accurate and has less influence on perception of auditory speech than with positive grey scale faces [9], [10]. However, visual speech presented in negative was still highly recognisable and influenced auditory speech recognition. These findings suggest that the luminance boundaries that are preserved in photo-negative could be important for perception of visual and audiovisual speech.

1. EFFECTS OF LINE DRAWN FACES ON VISUAL SPEECH PERCEPTION

Line drawing displays should preserve the primary boundaries between luminance differences that specify edges and contours in grey scale faces [11]. However, patterns of luminance across the face (for example, texture and shading) are removed [12]. Findings with static facial images show that presenting faces as line drawings considerably impairs recognition [11], [12], [13]. For example, Davies et al.[13] reported that subjects recognised the identity of line drawings of

[1] Maxine McCotter is now at the Department of Psychology, University of Stirling, Scotland.

familiar faces with 47% accuracy, compared to 90% accuracy with photographs of the same faces. Nevertheless, the outlines of the moving jaw, chin, cheeks, lips, teeth and tongue may provide sufficient information for visual and audiovisual speech perception. Preliminary findings with synthetic talking faces suggest that the facial contours and edges of the mouth are important for the perception of visual speech and improve recognition of congruent auditory speech [3], [14].

In order to investigate the role of edge based information and contours in the perception of visual and audiovisual speech, the present study presented talking faces in colour, grey scale and as line drawing displays. If basic luminance boundaries play a role in perception of visual and audiovisual speech, visual speech presented in line drawn faces should be highly recognisable and influence perception of visual and audiovisual speech. If this information is not important, visual speech presented in line drawn faces should have little influence on visual and audiovisual speech perception. If the distribution of luminance across the face is important in perception of visual speech, visual speech presented in colour and grey scale faces should be perceived more accurately and have a greater effect on auditory speech recognition than line drawn faces. If this information is not important, performance with colour, grey scale and line drawn faces should be the same. The present study used line drawn facial images where edges were digitally extracted from a natural talking face rather than synthesised, which should preserve natural facial characteristics (e.g., luminance boundaries, natural facial movement).

Experiments 1 and 2 investigated the effects of the three display types on auditory speech perception using unimodal auditory speech stimuli, unimodal visual speech stimuli, congruent audiovisual stimuli in which visual speech matched the auditory signal and incongruent (McGurk) audiovisual stimuli. Experiment 1 investigated effects of display type using auditory signals presented in quiet conditions. Experiment 2 investigated effects of display type using auditory signals of a lower intelligibility than in Experiment 1 in order to provide the opportunity for increased sensitivity to visual speech influences across the three display types. Using less audible speech sounds should lower overall levels of performance in congruent and auditory conditions sufficiently to reveal effects of display type on performance with congruent audiovisual stimuli, should they exist. Also, a great deal of research suggests that the influence of visual speech on auditory speech recognition increases when the signal-to-noise ratio of the auditory signal decreases [1], [7].

2. METHODS

Participants. Twenty-four adult native English speakers took part in 1 session lasting 1 hour 45 minutes. All participants reported normal or corrected to normal vision and good hearing. Twenty-four new participants from the same population as Experiment 1 took part in Experiment 2.

Stimuli. The face and voice of a 20-year- old female was recorded onto digital videotape using a digital video camera. The talker sat 1m from the camera and recordings were made of the talker saying 6 syllables. Each word was articulated normally with no artificial emphasis on articulation. The talker's face was fully illuminated and recorded against a dark background with only the face and neck visible. One example of each syllable was selected from the recordings and captured onto computer. The digitised clips were edited to produce the following stimuli: *Auditory speech*: auditory /ba/, /ga/, /ka/, /pa/, /ma/ and /ta/, all presented with a static face corresponding to each display type. *Visual speech:* visual /ba/, /ga/, /ka/, /pa/, /ma/ and /ta/ all with no auditory signal. *Congruent audiovisual speech*: auditory /ba/, /ga/, /ka/, /pa/, /ma/ and /ta/ paired with their corresponding auditory signal. *Incongruent audiovisual speech*: auditory /ba/ paired with visual /ga/, auditory /ga/ paired with visual /ba/, auditory /pa/ paired with visual /ka/, auditory /ka/ paired with visual /pa/, auditory /ma/ paired with visual /ta/, auditory /ta/ paired with visual /ma/. These combinations were pilot tested and found to produce powerful McGurk effects.

Grey scale stimuli were created by applying a grey scale filter to the digitised colour clips using Radius

Edit D.V. Line drawn stimuli were created by applying a 'find edges' filter to the grey scale stimuli using Adobe Aftereffects. The clips were recorded onto S-VHS tape, 1 tape for each display type. Tapes were played back on a high resolution visual display screen with auditory signals presented via 2 adjacent loudspeakers at a sound level of approximately 55dB. In Experiment 2, the same stimuli were presented in a background of continuous white noise at a sound level of 55dB (producing a signal to noise ratio of 0 dB). The monitor settings for colour faces were adjusted to produce naturalistic levels of colour, and contrast and brightness settings were the same for colour, grey scale and line drawing faces. Colour and grey scale images were matched for luminance using a Cambridge Research Systems OptiCAL photometer. For visual, congruent and incongruent stimuli, the face remained static for 2 seconds before onset of articulation, with each trial lasting 4 seconds, followed by a 3 second blank. For auditory speech stimuli, the face remained static throughout, with the onset of the auditory signal 2 seconds after visual stimulus onset.

Design. Each tape comprised 8 cycles of 24 stimuli, using a different random order for each cycle. Each tape started with a display of all 24 stimuli shown as practice items. Each participant saw all 3 display conditions. Order of presentation condition was counterbalanced across participants.

Procedure. Each participant was seated at a table approximately 1m in front of the display screen with a booklet placed on the table in front of them containing twelve response alternatives ("ba", "bga", "da", "ga", "ka", "ma", "la", "na", "pa" "pka", "tha", "ta") for each trial. Pre-testing had established that these twelve responses constituted more than 82% of participant's perceptions of the stimuli used in this experiment. Participants were instructed to look at the screen and listen throughout the experiment, and to make their responses by marking the syllable they heard (with the emphasis in the instructions being on heard) on their response sheet. When presented with unimodal visual speech trials, participants were instructed to report the syllable they thought was being articulated. Examples of stimuli were presented as practice items at the start of each display condition. Participants had a short break between each of the three display conditions.

3. RESULTS

A mixed design ANOVA with 1 between subject factor of order of presentation (e.g., colour, grey scale, line drawn; line drawn, grey scale, colour, etc.), and within subjects factors of speech condition (auditory, visual, congruent, incongruent), display type (colour, grey scale, line drawn) revealed no main effect or interaction involving order of presentation (all *p*s>.3). All subsequent analyses proceeded without presentation order as a factor.

Experiment 1. Mean percent correct responses to auditory, visual, congruent and incongruent speech stimuli are shown in Figure 1.

Auditory speech: A one way within subjects ANOVA with a factor of display type revealed that auditory stimuli were essentially equally intelligible in each display type (all *p*s<.5).

Visual speech: A one way ANOVA revealed a significant effect of display type, [$F(2,46)=4.41$, $p<.0178$]. Newman-Keuls tests showed that performance with colour and grey scale faces was more accurate than line drawn faces (both *p*s<.02). There were no differences between performance with colour and grey scale faces (*p*>.7).

Congruent audiovisual speech: A one way ANOVA showed no significant main effect (*p*>. 09). The data obtained with congruent stimuli were compared with those obtained with the same auditory signals in the auditory speech condition. An ANOVA with factors speech condition (auditory, congruent) and display type showed no significant main effects or interactions (all *p*s>.1).

Incongruent audiovisual speech: A one way ANOVA revealed a significant effect of display type, [$F(2,46)=7.31$, $p<.0018$]. Newman-Keuls tests showed no difference in performance accuracy between colour and grey scale faces, (*p*>.5) and higher accuracy for line drawing faces (i.e., fewer McGurk responses were made; both *p*s<.005).

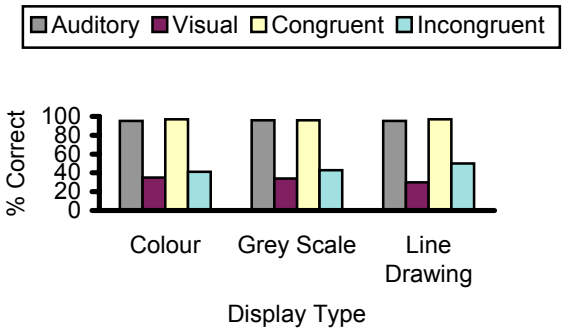
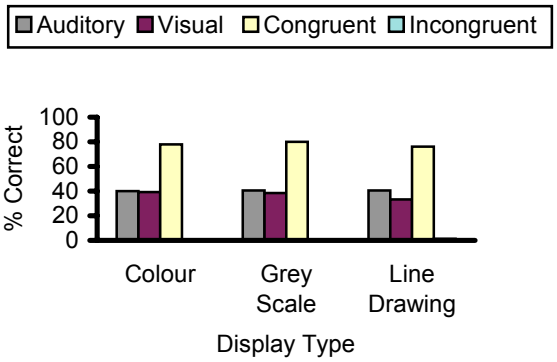


Figure 1A Mean percentage of auditory and visual stimuli and the auditory component of congruent and incongruent audiovisual stimuli correctly identified in Experiment 1 for each display type.

The data obtained with incongruent audiovisual speech were compared with that obtained with the same auditory signals in the auditory speech condition. An ANOVA with factors speech condition (auditory, incongruent) and display type revealed significant effects of speech condition, [$F(1,23)=78.34$, $p<.0001$], display type, [$F(2,46)=5.10$, $p<.0010$], and an interaction between speech condition and display type, [$F(2,46)=7.37$, $p<.0017$]. Newman-Keuls tests showed fewer correct responses in all display types and stimulus types for incongruent audiovisual stimuli when compared with the auditory speech condition (all ps<.0002). This deficit was smaller for line drawing faces than for colour and grey scale faces (both ps<.0001).

Experiment 2. Mean percent correct responses to auditory, visual, congruent and incongruent speech stimuli presented in auditory noise are shown in Figure 2. *Auditory speech*: A one way ANOVA with factors display type (colour, grey scale, line drawing) revealed that auditory stimuli were equally intelligible in each display type [$F(2,46)=1.53$, $p>.2$].
Visual speech: A one way ANOVA revealed a significant effect of display type, [$F(2,46)=8.17$, $p<.0009$]. Newman-Keuls tests showed that identification of visual speech presented in colour and grey scale faces was more accurate than visual speech presented in line drawing faces (both ps<.0004).



[2] **Figure 2 Mean percentage of auditory and visual stimuli and the auditory component of congruent and incongruent audiovisual stimuli correctly identified in Experiment 2 (noise condition) for each display type.**

There were no differences between performance with colour and grey scale faces ($p>.5$).
Congruent audiovisual speech: A one way ANOVA revealed no significant main effects (all ps>1). The data obtained with congruent stimuli were compared with those obtained with the same auditory signals in the auditory speech condition. An ANOVA with factors speech condition (auditory, congruent) and display type showed an effect of speech condition, [$F(1,23)=374.2$, $p<.0001$]. Newman-Keuls tests showed that congruent audiovisual speech improved perception of auditory speech ($p<.0002$).
Incongruent audiovisual speech: A one way ANOVA revealed no significant main effect

[2]Due to low levels of performance with auditory stimuli, accuracies with incongruent audiovisual speech do not appear on the graph axes. These performance accuracies were 0.4.%, 0.7% and 1.2% correct for colour, grey scale, and line drawn faces respectively.

($p>.09$)[3]. The data obtained with incongruent audiovisual speech were compared with that obtained with the same auditory signals in the auditory speech condition. An ANOVA with factors speech condition (auditory, congruent), display type and stimulus type revealed a significant effect of speech condition [$F(1,23)=349.67$, $p<.0001$]. Newman-Keuls tests showed fewer correct responses for incongruent audiovisual stimuli when compared with the auditory speech condition (all ps$<.0002$)

4. DISCUSSION

The main findings of Experiments 1 and 2 are as follows. Visual speech presented in line drawn faces was highly recognisable, although performance with line drawn faces was marginally (but significantly) less accurate than performance with colour and grey scale faces. The levels of visual influence obtained with colour, grey scale and line drawn faces were equally effective for the perception of congruent auditory speech. Also, visual speech presented in line drawn faces was highly influential for perception of incongruent auditory speech stimuli (i.e., these images produced significant McGurk effects). However, Experiment 1 reported that visual speech presented in line drawn faces was less influential on perception of incongruent auditory speech than colour and grey scale faces.

Findings with colour and grey scale faces replicate those of Jordan et al., [6] showing no differences in performance between colour and grey scale faces. This adds weight to the notion that luminance, rather than colour, plays an important role in visual speech perception.

In sum, line drawn faces exerted considerable influence in all speech conditions, often providing levels of visual influence on perception of auditory speech equivalent to that provided by colour and grey scale faces. These findings suggest that information available in these displays plays a crucial role in visual and audiovisual speech perception. Line drawn faces should retain luminance boundaries, which specify contour and edge based information. These luminance boundaries could specify the outlines of the jaw, chin, cheeks, lips, teeth and tongue, providing information for visual speech perception. [3,] [4], [5]. For example, the visible contour of the tongue in the mouth may help to specify the tip of the tongue as it protrudes during speech. This information could provide cues to tongue position, which may be important in the perception of visual speech [4]. Luminance boundaries can also help to specify the parameters of extra-oral features such as eyes, jaw, chin and cheeks, which may contribute to recognition of visual and audiovisual speech [15].

The slight performance deficit with visual speech presented in line drawn faces suggests that the distribution of luminance across the face plays some role in perception of visual speech. Subtle variations in dark and light provided by the distribution of luminance may convey skin pigmentation and texture. For example, the teeth are smoother than the rest of the mouth region, and the tongue has a characteristic rough texture compared to the lips. These subtle luminance differences may emphasise distinctions between the lips, mouth, tongue, cheeks and chin, which in turn, could contribute to the perception of visual speech [3], [4], [5].

The present findings suggest that congruent audiovisual speech can be processed effectively via basic luminance boundaries present in line drawn faces without the need for additional, more complex, luminance information. However, the distribution of luminance across the face appears to play a role in the processing of unimodal visual speech. Why might this be? Perception of visual

[3] Accuracies with incongruent auditory speech in Experiment 2 were particularly low, (e.g., 1% correct compared with 45% correct with incongruent audiovisual stimuli presented in Experiment 1). Approximately 99% of responses reported with incongruent speech stimuli in Experiment 2 were McGurk responses. Thus, the lack of an effect of display type on perception of these particular incongruent auditory speech stimuli should be interpreted with caution, as performance differences between stimuli presented in the three display types may have been suppressed.

speech may require specific visual information to distinguish between viseme groups (e.g., /b/ and /g/) and also within viseme groups (e.g., /b/ and /p/, [16], [17]). Thus, additional visual information provided by the distribution of luminance may aid these within viseme group distinctions. For example, the tongue shades gradually into the surrounding tissues, with the darker regions of the tongue inside the mouth contrasting with the lighter areas on the tip of the tongue as it protrudes. This shading information could specify the position of the tongue in the mouth, which may in turn emphasise the physical differences in the articulations formed near the back of the mouth (e.g., /g/ and /k/, [16]). From a theoretical perspective, the present findings suggest a hierarchy in the contribution of luminance to visual and audiovisual speech recognition. When only the visual input is provided, the distribution of luminance across the face increases the information value of visible articulations necessary for identification of the speech sound. When audio and visual inputs are provided, basic luminance boundaries appear to be sufficient.

5. REFERENCES

1. MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 12,* 131-141.
2. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.
3. Brooke, N.M., & Summerfield, Q. (1983). Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics, 11,* 63-76.
4. Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, Massachusetts: MIT press.
5. Summerfield, Q., MacLeod, A., McGrath, M., & Brooks, M. (1989). Lips, teeth, and the benefits of lipreading. In A.W. Young & H.D. Ellis (Eds.), *Handbook of Research on Face Processing* (pp. 223-233). Holland: Elsevier.
6. Jordan, T.R., McCotter, M.V., & Thomas, S.M. (2000). Visual and audiovisual speech perception with color and gray scale facial images. *Perception & Psychophysics, 62,* 1394-1404.
7. Jordan, T.R., & Sergeant, P.C. (1998). Effects of Facial Image Size on Visual and Audio Visual Speech Recognition. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by Eye Part 2: The Psychology of Speechreading and Audiovisual Speech (*pp. 155-176). London: Taylor & Francis Press.
8. Rosenblum, L.D., & Saldaña , H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 318-331.
9. Kanzaki, R., & Campbell, R. (1999). Effect of facial brightness reversal on visual and audiovisual speech perception. *Paper presented at the AVSP conference, 1999.*
10. McCotter , M.V., & Jordan, T.R. (2001). *Investigating the role of facial luminance in visual and audiovisual speech perception.* Manuscript under review.
11. Leder, H. (1999). Matching person identity from facial line drawings. *Perception, 28,* 1171-1175.
12. Bruce, V., Hanna, E., Dench, N., Healey, P., & Burton, M. (1992). The importance of 'mass' in line drawings of faces. *Applied Cognitive Psychology, 6,* 619-628.
13. Davies, G., Ellis, H., & Shepherd, J. (1978). Face recognition accuracy as a function of mode of representation. *Journal of Applied Psychology, 63, 180-187.*
14. Dalton, B., Kaucic, R., & Blake, A. (1996). Automatic speechreading using dynamic contours. In D.G. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series F; Computer and Systems sciences, Vol. 150, pp373-382). Berlin: Springer-Verlag.
15. Vatikiotis-Bateson, E., Eigsti, I.M., Yano, S., & Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60,* 926-940.
16. Jeffers, J. & Barley, M. (1971). *Speechreading* Springfield, Illinois: C.C. Thomas
17. Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr,C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130-145.