Journal of Experimental Psychology: Human Perception and Performance

Searching for a Talking Face: The Effect of Degrading the Auditory Signal

Paula C. Stacey, Thomas Murphy, Christian J. Sumner, Pádraig T. Kitterick, and Katherine L. Roberts

Online First Publication, October 20, 2014. http://dx.doi.org/10.1037/a0038220

CITATION

Stacey, P. C., Murphy, T., Sumner, C. J., Kitterick, P. T., & Roberts, K. L. (2014). Searching for a Talking Face: The Effect of Degrading the Auditory Signal. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 2106–2111.

Running head: SEARCHING FOR A TALKING FACE

Searching for a talking face:

The effect of degrading the auditory signal

Paula C. Stacey[1], Thomas Murphy[1], Christian J. Sumner[2], Pádraig T. Kitterick[3],

and Katherine L. Roberts[4]

[1] Division of Psychology, Nottingham Trent University, Nottingham, UK

[2] MRC Institute of Hearing Research, University Park, Nottingham, UK

[3] NIHR Nottingham Hearing Biomedical Research Unit, Ropewalk House, Nottingham, UK

[4] Department of Psychology, University of Warwick, Coventry, UK

Abstract

Previous research (e.g. McGurk and MacDonald, 1976) suggests that faces and voices are bound automatically, but recent evidence suggests that attention is involved in a task of searching for a talking face (Alsius and Soto-Faraco, 2011). We hypothesised that the processing demands of the stimuli may affect the amount of attentional resources required, and investigated what effect degrading the auditory stimulus had on the time taken to locate a talking face. Twenty participants were presented with between 2 and 4 faces articulating different sentences, and had to decide which of these faces matched the sentence that they heard. The results showed that in the least demanding auditory condition (clear speech in quiet), search times did not significantly increase when the number of faces increased. However, when speech was presented in background noise or was processed to simulate the information provided by a cochlear implant, search times increased as the number of faces increased. Thus, it seems that the amount of attentional resources required vary according to the processing demands of the auditory stimuli, and when processing load is increased then faces need to be individually attended to in order to complete the task. Based on these results we would expect cochlear-implant users to find the task of locating a talking face more attentionally demanding than normal hearing listeners.

Searching for a talking face: The effect of degrading the auditory signal

Combining auditory and visual information is an important perceptual task. Understanding speech in background noise can be difficult for both normal hearing (Sumby & Pollack, 1954) and hearing impaired listeners (Davis, 1989; Thibodeau, 2004), but seeing the face of the talker helps both groups of people (MacCleod & Summerfield, 1990; Larsby, Hällgren, Lyxell, & Arlinger, 2005). In order to benefit from "visual speech" information, people need to locate the talker of interest. Whether combining faces and voices is automatic or requires selective attention has been debated recently.

Experiments using the McGurk effect (McGurk & MacDonald, 1976) have suggested that faces and voices are bound without the need for selective attention (e.g. Massaro, 1987; Walker, Bruce, & O'Malley, 1995; Soto-Faraco & Alsius, 2009). Additionally, Van der Burg, Olivers, Bronkhorst, and Theeuwes (2008) found that non-spatial auditory signals can guide attention towards synchronised visual events, and allow visual targets to 'pop out' in the scene. The finding that synchronous visual and audio events are perceptually grouped is supported by Roseboom, Nishida, Fujisaki and Arnold (2011) who found that the ability to identify a synchronous stream of audio-visual speech was enhanced by the presence of simultaneous streams of asynchronous visual speech. Event-related potential (ERP) studies support claims that these perceptual effects reflect early, and potentially pre-attentive, multisensory integration of auditory and visual stimuli (Colin et al. 2002, Van der Burg,Talsma, Olivers, & Theeuwes, 2011).

More recently however, research has suggested that attention may play a role in combining face and voice information. Increasing cognitive load by adding a secondary task can decrease the magnitude of the McGurk effect (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007). Additionally, selective attention appears to be necessary to bind faces and voices when there are several faces present. Alsius and Soto-Faraco (2011) presented participants with 4 talking faces, each of which was articulating a different sentence. They used precues to direct participants' attention to between 2 and 4 of these faces, and found that the time taken to locate the correct face increased as the number of cued locations increased. In summary, it seems that although the integration of auditory and visual stimuli might be automatic, this may vary according to task demands.

One way in which the task of combining face and voice information might become more attentionally demanding is if the auditory signal is degraded, as is the case for people who have cochlear implants. A cochlear implant is an electronic device which restores partial hearing to people who are profoundly deaf. By stimulating the auditory nerve directly via an electrode array which has been surgically implanted into the inner ear (the cochlea), cochlear implantation restores the audibility of sounds (Bond *et al*., 2009). However, the signals that implantees receive are degraded spectrally (they receive fewer channel of information) and temporally (limited to slow fluctuations in amplitude over time). These processing limitations are particularly detrimental to the ability to understand speech in noise (Turner, Gantz, Vidal, Behrens, & Henry, 2004).

The current study investigates whether the demands of the auditory stimuli affect the ability to locate a talking face. The auditory speech signal was distorted by: (1) processing with a sine-wave vocoder which simulated the distortions in speech faced by cochlear-implant users; and (2) adding background noise. We expected that degrading the speech signal would lead to increases in the time taken to locate the matching talking face as the number of faces increased, and hypothesised that differences between previous studies could be accounted for by the processing demands of the stimuli.

Method

Design

A 3 (Number of Faces: 2, 3, or 4; within) x 2 (Noise: Quiet or Noisy; within) x 2 (Speech Type: Clear or Vocoded; between) mixed design was used. The dependent variable was the time (in milliseconds) taken to select the face which matched the spoken sentence.

Participants

Twenty-four students (14 male, mean age 20.3 years) from the Nottingham Trent University took part. All reported having normal hearing, normal or corrected-to-normal vision, and spoke English as their first language. Ethical approval was granted by the Nottingham Trent University.

Apparatus and Materials

Audiovisual recordings

The materials used were 90 IEEE (IEEE, 1969) sentences recorded audiovisually and spoken by a single male talker with a British accent. An example sentence from this corpus is “The slang name for all alcohol is booze.” The auditory speech was recorded at a sample rate of 44100 Hz and the visual speech at rate of 25 frames per second. Each sentence was approximately 3 seconds long.

Signal processing

Matlab (The Mathworks, Nantick, US) was used to first embed the sentences in background noise. Multi-talker babble was added at a signal-to-noise ratio (SNR) of -4 dB for the clear speech condition and +3 for the vocoded speech condition. These different SNRs were selected as they lead to 80% correct audio-only speech perception performance (unpublished data).

The mixed signals were then processed using a sine-wave vocoder. Signals were band-pass filtered into 8 adjacent frequency bands, spaced equally on an ‘equivalent rectangular bandwidth’ ($ERB_{N,}$ Glasberg and Moore, 1990) frequency scale between 100Hz and 8kHz. In natural speech conditions, the auditory stimuli were constructed by summing the output of the 8 band-pass filters. In conditions where the speech was vocoded, the Hilbert transform was used to modulate a pure tone at the centre frequency of the respective filter. The sine waves were then summed to form the vocoded signal.

Apparatus

Stimuli were presented using EPrime over Seinheisser HD280pro headphones. Custom built hardware provided by the MRC Institute of Hearing Research was used to perform digital-to-analogue conversion and amplification for presentation over headphones at a calibrated sound pressure level (SPL). Stimuli were presented on a computer screen measuring 44.5 x 25.4 cm.

Procedure

The experiment took place in a quiet room. On each trial 2, 3, or 4 talking faces were presented on the computer screen, each articulating a different IEEE sentence. Each mpg file was presented 17cm high by 10cm tall, and participants were seated approximately 50cm away from the monitor (see Figure 1 for an illustration of how the faces were presented on screen in each condition). Auditory stimuli were presented at an average sound level of 70 dB SPL. The auditory sentence that corresponded to one of the talking faces was presented at the same time as the visual stimuli, and participants were asked to use the computer mouse to select the talking face that matched the auditory sentence. They were asked to respond as quickly but as accurately as possible. Between each trial, participants were instructed to fixate a centrally-presented cross which was presented for one second. The sentences used for the audiovisually-incongruent distractor faces were selected randomly (with the exclusion of the target sentence) from the database of sentences, with the restriction that each sentence was used an equal number of times throughout the experiment.

----Insert Figure 1----

Fifteen practice trials were administered before the experiment. These consisted of 5 faces in each of the 2, 3, and 4 face conditions, which were presented in blocks in a counterbalanced order. The experiment comprised 90 trials: 30 in each of the 2, 3, and 4 face conditions. The number of faces in each block of trials was counterbalanced across conditions.

Analyses

Response times for each individual participant were screened, and any data points more than 2 standard deviations from the mean were removed (see Ratcliff, 1993). Data from correct trials only were entered into the response time analysis, so we required participants to score over 80% correct in each condition to be included.

Results

Three participants were excluded for having accuracy levels less than 80% in one or more conditions, and data storage for one participant failed. Therefore, the following analyses are based on 9 participants in the Clear condition and 11 participants in the Vocoded condition (the significance of all main effects and interactions was unaffected by excluding these 3 participants). Overall accuracy levels were high, with participants responding correctly on 93.94% of trials (standard deviation 6.86). The overall average response time was 2439 milliseconds (ms; standard deviation 556 ms), and average response times for all remaining participants fell within 2 standard deviations of the mean.

Response time analysis

Figure 2 shows average response times when there were 2, 3, or 4 faces on screen for participants in Clear or Vocoded speech conditions, with or without noise. The overall pattern suggests that the impact of additional faces on screen is larger when the processing demands of stimuli are increased through vocoding or by adding background noise. A 3x2x2 mixed ANOVA revealed significant main effects of Number of Faces ($F_{2, 36} = 6.04$, $MS_e = 105185.04$, $p=0.005$, $\eta_p^2= 0.25$) and Noise ($F_{1, 18} = 30.94$, $MS_e = 24434.66$, $p<0.001$, $\eta_p^2= 0.63$), and the main effect of Speech Type just failed to reach significance ($F_{1, 18} = 4.24$, $MS_e = 1368971.68$, $p=0.054$, $\eta_p^2= 0.19$). There was additionally a significant three-way interaction between Number of Faces, Noise, and Speech Type ($F_{2, 36} = 5.77$, $MS_e = 26357.50$, $p=0.007$, $\eta_p^2= 0.24$).

----Insert Figure 2----

Two separate 3 (Number of faces) x 2 (Noise) ANOVAs were carried out on clear and vocoded speech respectively to follow up this three-way interaction (Table 1). These analyses revealed that there was a significant 3 x 2 interaction between Number of Faces and Noise for the Clear Speech condition, but not for the Vocoded condition. In the clear-quiet condition, search times did not significantly increase according to number of faces on screen, but when speech was in background noise search times did increase with increasing number of faces (Table 2). In contrast, for the Vocoded conditions the effect of increasing the number of faces occurred irrespective of whether the speech was presented in quiet or in background noise (Figure 2).

----Insert Table 1----

----Insert Table 2----

Accuracy

For the accuracy data there was a significant main effect of Number of Faces ($F_{2, 36} = 12.97$, $MS_e = 35.02$, $p<0.001$, $\eta_p^2= 0.42$; Figure 3). Accuracy was poorer when there were 4 faces on screen (average = 90.17% correct, standard deviation = 7.84) compared with when there were 2 (average = 96.83% correct, standard deviation = 5.44; $t_{39} = 4.16$, $p<0.001$) or 3 faces (average = 94.83% correct, standard deviation = 5.33; $t_{39} = 3.39$, $p=0.002$) present. No other main effects of interactions reached significance.

----Insert Figure 3----

Discussion

The study investigated the effects of degrading the auditory signal on the time taken to locate a talking face. When speech was at its least degraded (in the clear quiet condition), there was no significant effect of increasing the number of faces in the search array. When the speech signal was degraded however, either through the addition of background noise or through reducing the spectral and temporal resolution of speech in the vocoder conditions, search times increased with increasing number of faces on screen. These results suggest that the amount of attentional resources required vary according to the processing demands of the auditory stimuli, and when processing load is increased then faces need to be individually attended to in order to complete the task.

The findings of this experiment support the conclusions drawn by Navarra Alsius, Soto-Faraco and Spence (2010) and Spence and Deroy (2013), who argued that the automaticity of audiovisual integration depends on the specific demands of a given task. With a low processing load, the results are consistent with the original McGurk findings. However, the results are also consistent with studies which have shown that cognitive load affects the ability to combine face and voice information (Alsius *et al*., 2005; 2007). The results are somewhat contradictory to Alsius and Soto-Faraco (2011), since they used natural unprocessed speech and found that search times increased with the number of cued faces. Differences in the methodology could partly explain these differences; while Alsius and Soto-Faraco (2011) always displayed 4 faces and cued between 2 and 4 target locations, we only display potential target faces. The amount of visual crowding therefore varied in our experiment, while it did not in the study by Alsius and Soto-Faraco (2011). However, we would expect our procedure to lead to more marked increases in search times as the number of faces increased, rather than flattening the response curves for the 'Clear-Quiet' condition. It is also possible the auditory intelligibility of the talkers used varied across studies. The talker used in the current experiment has been shown to be highly intelligible in auditory-only conditions even if the speech is degraded (Stacey & Summerfield, 2007).

Another difference between our study and Alsius and Soto-Faraco (2011) is that they faded in and out their auditory stimuli while the faces were already moving, to avoid the possibility that abrupt onsets could provide a cue (as shown by Van der Burg, Cass, Olivers, Theeuwes, & Alais, 2010), whereas we did not. However, there are a number of reasons to suggest that abrupt onsets are not behind the differences we find between our groups. First, the onsets for the 'Vocoded-Quiet' condition are just as abrupt as for the 'Clear-Quiet' condition. Second, we found no relationship between the onset times of the stimuli we used against reaction times. Third, we expect responses to be quicker if onsets were providing a powerful cue. Potentially collecting responses using a mouse-click was problematic since the distance of faces from one another varies across the 2, 3, and 4 face conditions. However, while the response procedure may have obscured some differences between the number-of-faces conditions, when all four speech type conditions are taken together robust differences remain evident.

Visual search times were longer overall in the vocoded conditions than in the clear speech conditions, despite overall intelligibility being similar. Sine-wave vocoding degrades the speech signal by providing fewer separate channels of information, and removing small amplitude fluctuations over time. We cannot say which of these degradations was most important here. However, we can infer that users of cochlear implants will both take longer to find a talking face in a crowd, and will find the addition of more people more attentionally demanding. These are important issues for cochlear-implant users because they find listening to speech in noisy environments difficult (Turner *et al*., 2004), and visual speech information has been shown to improve performance (Kaiser, Kirk, Lachs, & Pisoni, 2003; Grant, Walden, & Seitz, 1998). Previous research suggests that computer-based auditory training can improve speech perception amongst people with implants (Fu, Galvin, Wang, & Nogaki, 2005; Stacey *et al*., 2010; Ingvalson, Lee, Fiebig, & Wong, 2013), and it is also possible they would benefit from training to locate speakers in a multi-talker array.

To conclude, this study suggests that the amount of attentional resources required to locate a talking face varies according to the processing demands of the auditory stimuli. These results suggest that users of cochlear implants will find the task of locating a talking face in a multi-speaker scenario more difficult and more attentionally demanding than normal-hearing listeners.

References

Alsius, A., Navarra J., Campbell R., & Soto-Faraco S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15,* 839–843. doi: 10.1016/j.cub.2005.03.046

Alsius, A., Navarra J., & Soto-Faraco S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research, 183*, 399–404. doi: 10.1007/s00221-007-1110-1

Alsius, A., & Soto-Faraco S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. *Experimental Brain Research. 213*, 175-183. doi: 10.1007/s00221-011-2624-0

Bond, M., Mealing, S., Anderson, R., Elston, J., Weiner, G., Taylor, R.S., Hoyle, M., Liu, Z., Price, A., & Stein, K. (2009). The effectiveness and cost-effectiveness of cochlear implants for severe to profound deafness in children and adults: a systematic review and economic model. *Health Technology Assessment, 13*, 1-330. doi: 10.3310/hta13440

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., &. Delenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology, 113*, 495–506. doi: 10.1016/S1388-2457(02)00024-X.

Davis, A.C. (1989). The prevalence of hearing impairment and reported hearing disability among adults in Great Britain. *International Journal of Epidemiology, 18,* 911-917. doi: 10.1093/ije/18.4.911.

Fu, Q-J., Galvin, J.J., Wang, X., & Nogaki, G. (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoustic Research Letters Online, 6,* 106-111. doi: 10.1121/1.1898345.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103–138.

Grant, K.W., Walden, B.E., & Seitz, P.F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103,* 2677–2450.

IEEE. (1969). *IEEE Recommended Practice for Speech Quality Measurements*. New York: Institute for Electrical and Electronic Engineers.

Ingvalson, E.M., Lee, B., Fiebig, P., & Wong, P.C.M. (2013). The effects of short-term computerized speech-in-noise training on postlingually deafened adult cochlear implant recipients. *Journal of Speech Language and Hearing Research, 56,* 81-88. doi:10.1044/1092-4388.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46,* 390–404. doi:10.1044/1092-4388.

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology, 44,* 131-143. DOI: http://dx.doi.org/10.1080/14992020500057244

MacLeod, A., & Summerfield A.Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24*, 29-43.

Massaro, D.W. (1987). Speech perception by ear and eye. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The psychology of lip-reading* (pp53-83). Hillsdale, NJ: Erlbaum.

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

Navarra, J., Alsius A., Soto-Faraco S., & Spence C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion, 11*, 4–11. doi: 10.1016/j.inffus.2009.04.001

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin 114,* 510-532.

Roseboom, W., Fujisaki, W., Nishida, S. & Arnold, D.H. (2011). Audio-visual speech timing sensitivity is enhanced in cluttered conditions. *PLoS One, 6,* e18309. doi: 10.1371/journal.pone.0018309.

Soto-Faraco, S., & Alsius A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology. Human Perception and Performance, 35*, 580–587. doi: 10.1037/a0013483.

Spence, C., & Deroy, O. (2013) How automatic are crossmodal correspondences? *Consciousness and Cognition, 22*, 245-260. doi: http://dx.doi.org/10.1016/j.concog.2012.12.006.

Stacey, P.C., & Summerfield, A.Q. (2007). Effectiveness of computer-based auditory training in improving the ability of normally-hearing listeners to understand spectrally-distorted speech. *Journal of the Acoustical Society of America, 121*, 2923-35. doi: 10.1121/1.2713668

Stacey, P.C., Raine, C.H, O'Donoghue, G.M., Tapper, L., Twomey, T., & Summerfield, A.Q. (2010). Effectiveness of computer-based auditory training for adult users of cochlear implants. *International Journal of Audiology, 49*, 347-356. doi: 10.3109/14992020903397838.

Sumby, W., Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Thibodeau, L.M. (2004). Plotting beyond the audiogram to the TELEGRAM, a new assesment tool. *Hearing Journal, 57*, 46-51.

Turner, C.W., Gantz, B.J., Vidal, C. Behrens, A., & Henry, B.A. (2004). Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing. *Journal of the Acoustical Society of America, 115*, 1729–1735. doi: http://dx.doi.org/10.1121/1.1687425

Van der Burg, E., Cass, J., Olivers, C.N., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS One, 5,* e10664. doi: 10.1371/journal.pone.0010664.

Van der Burg, E., Olivers, C.N., Bronkhorst, A.W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology. Human Perception and Performance, 34,* 1053-1065. doi: 10.1037/0096-1523.34.5.1053.

Van der Burg, E., Talsma, D., Olivers, C.N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55*, 1208-1218. doi: 10.1016/j.neuroimage.2010.12.068.

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.

Table 1

Results from 3 (Number of Faces) x 2 (Noise) ANOVAs on visual search times for Clear and Vocoded speech. Significant results are shown in bold.

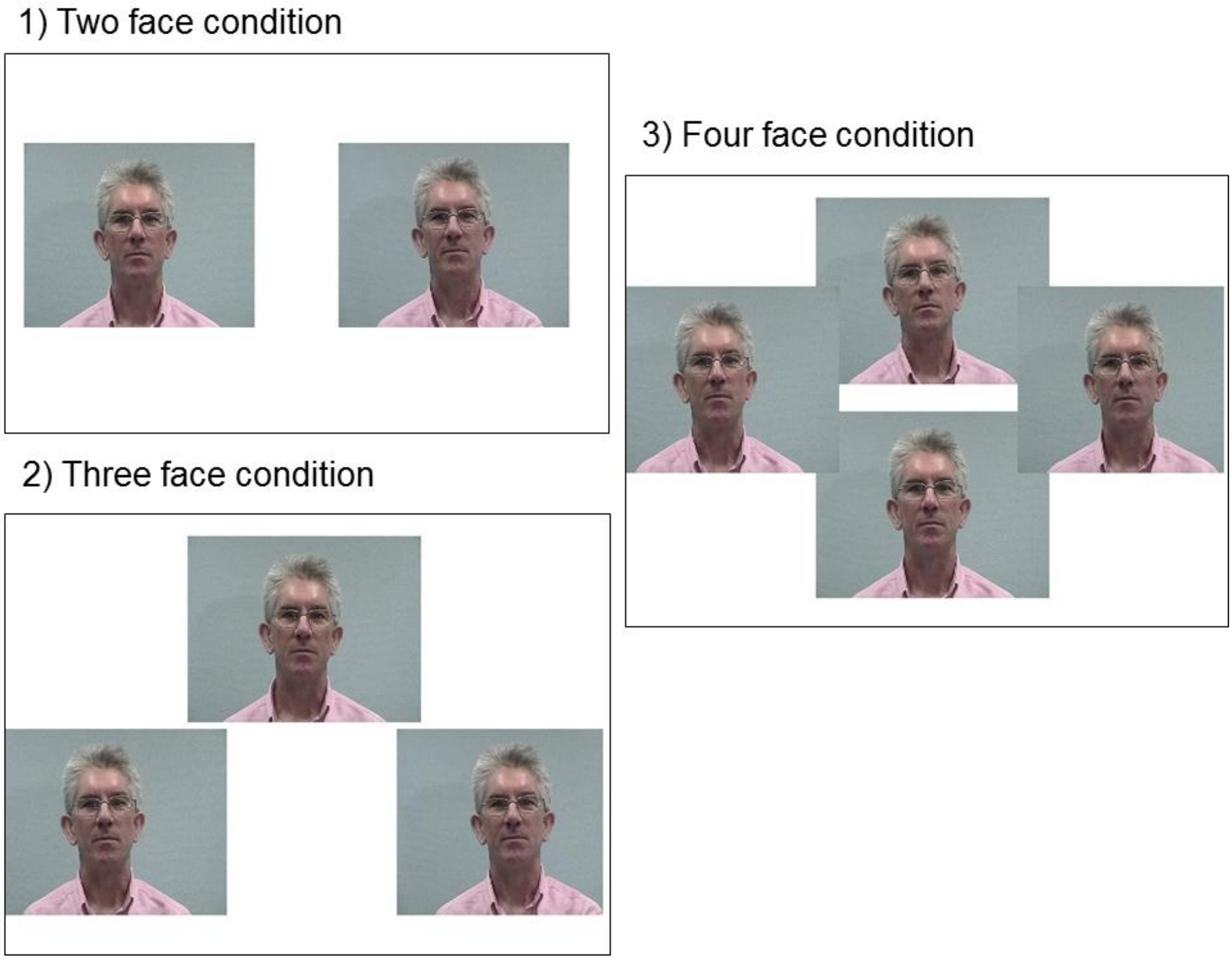
	Clear speech	**Vocoded speech**
Number of faces	$F_{2, 16} = 1.73$, $MS_e = 157280.56$, $p=0.21$, $\eta_p^2 = 0.18$	**$F_{2, 20} = 6.11$, $MS_e = 63508.63$, $p=0.008$, $\eta_p^2 = 0.38$**
Noise	**$F_{1, 8} = 21.80$, $MS_e = 24903.04$, $p=0.002$, $\eta_p^2 = 0.73$**	**$F_{1, 10} = 9.67$, $MS_e = 24059.96$, $p=0.011$, $\eta_p^2 = 0.49$**
Number of faces X Noise	**$F_{2, 16} = 4.50$, $MS_e = 37609.21$, $p=0.028$, $\eta_p^2 = 0.36$**	$F_{2, 20} = 2.29$, $MS_e = 17356.14$, $p=0.127$, $\eta_p^2 = 0.19$

Table 2

Results from one-way ANOVAs (Number of Faces) on visual search times in the Clear quiet and Clear Noise conditions. A Bonferroni correction for 2 comparisons has been applied.

	F	**df**	**MS_e**	**Sig**	**η_p^2**
Clear quiet	0.04	2, 16	135606.26	p=0.96	0.01
Clear noise	7.34	2, 16	59283.51	p=0.01	0.48

Figure 1. Arrangement of faces on screen in the two, three, and four face conditions.



1) Two face condition

2) Three face condition

3) Four face condition

Figure 2. Response times according to number of faces, presence of background noise, and speech type. Panel A shows data for Clear speech, and Panel B for Vocoded speech. Error bars indicate standard errors.

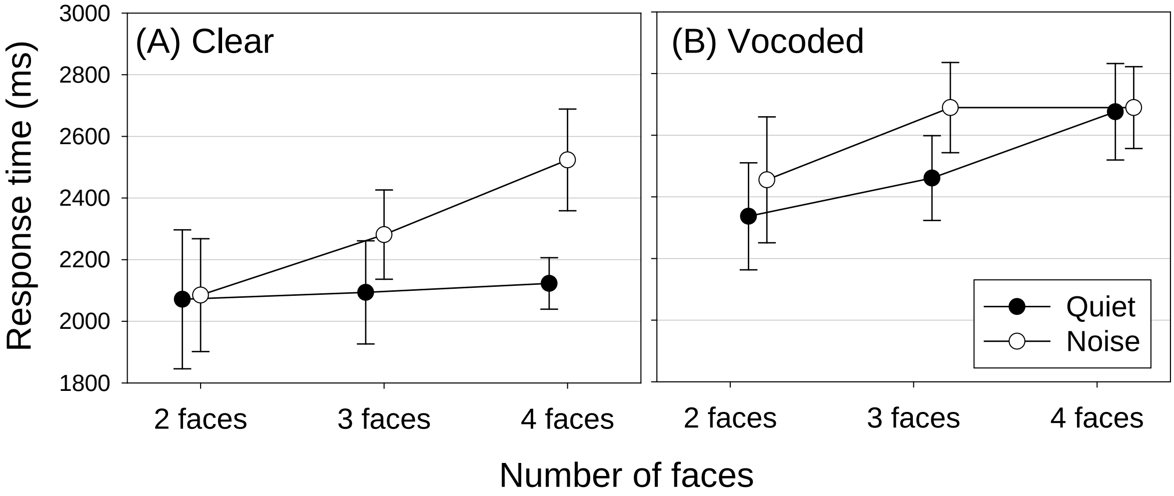
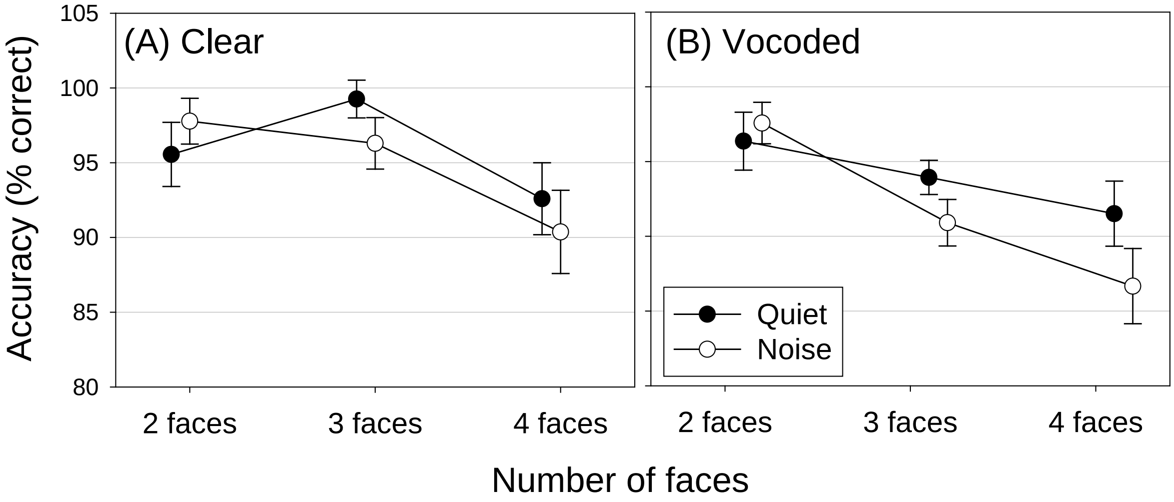


Figure 3. Accuracy data according to to number of faces, presence of background noise, and speech type. Panel A shows data for Clear speech, and Panel B for Vocoded speech. Error bars indicate standard errors.


(A) Clear
(B) Vocoded
Accuracy (% correct)
2 faces
3 faces
4 faces
Number of faces
● Quiet
○ Noise