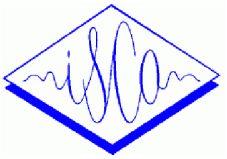
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions

Natalie Fecher & Dominic Watt

Department of Language and Linguistic Science, University of York, York, United Kingdom
{natalie.fecher|dominic.watt}@york.ac.uk

Abstract

The study presented in this paper investigates auditory-only and auditory-visual (AV) consonant recognition where the talker's face is obscured by various types of face-concealing garments and headgear. Observers' consonant identification performance across the various 'facewear' conditions was tested both in quiet listening conditions (Experiment 1), and when the speech stimuli were embedded in 8-talker babble noise (Experiment 2). Statistical analysis of the responses collected from 82 phonetically-untrained subjects (N = 43, quiet; N = 39, noise) revealed a significant AV effect in both experiments. However, the strength of the effect varied considerably as a function of facewear type. The findings are discussed in the context of previous research on AV speech perception, which aims to identify the facial regions that are particularly important for the extraction of visual speech cues.

Index Terms: auditory-visual speech perception, consonant identification, facial occlusion, forensic speech science

1. Introduction

It is by now well documented in the relevant literature that speech intelligibility is improved when both facial and auditory cues generated during speech production are available to the perceiver. Since the early work, e.g. [1, 2], it has been repeatedly demonstrated that the linguistic information derived from the acoustic signal and the visible speech gestures from the talker's articulating face combine synergistically into a coherent percept, which may be more complete than that obtained from either of the unimodal sources alone [3-13]. The multimodal nature of spoken language processing has been revealed by examining how missing speech cues in one channel can be recovered from the other channel, respectively, in cases where auditory or visual information is disrupted or lost from the signal. It has been shown, among other things, that in adverse listening conditions, e.g. when acoustic speech cues are absent or distorted by additive noise, listeners rely more heavily upon visual speech cues extracted from the lips, tongue, teeth, and/or cheeks [5, 9-11, 13]. Using ever more sophisticated video capture and post-processing techniques, researchers have focused further attention on the interaction of the two modalities when the image accompanying the auditory stimulus is partially or wholly obscured [2-4, 7, 8, 10, 12, 13]. The experimental techniques, the linguistic material tested, and the region of interest in the talker's face vary widely across studies, but one common aim has been to identify the facial areas which are most informative for the observer during AV speech processing. Overall, it was shown that the cognitive processes responsible for the perception of facial movement during AV speech perception are notably resistant to the loss of coarse (configural) facial information (e.g. due to facial inversion, or changes to distance, viewing angle, and colour) and fine facial detail (e.g. by modifying spatial resolution) [4, 8, 13].

For the purpose of determining the relative prominence of one region in a talker's face over another, the 'window technique' has proven particularly useful [3, 4, 7, 8, 12, 13]. Different facial areas are thereby systematically eliminated from view, and the effect on speech recognition is tested. This technique has been applied to studying the visual cues involved in both prosodic [3, 12] and segmental processing [4, 8]. However, despite the benefits this method offers, e.g. of investigating orofacial structures independently of one another, it has also faced criticism. The authors of [4] and [13] contend that selective masking of a talker's face may unintentionally induce unnatural viewing and attentional strategies, and may underestimate the role of holistic facial information during AV speech perception. For these reasons, they argue, researchers setting out to explore the extent to which subjects will tolerate loss of perceptual information that is brought about by facial occlusion should make use of more realistic occlusions in their studies [4]. They propose that "a natural system of visual and audiovisual speech perception is likely to develop to cope with everyday occlusions that do not obscure all of a face except for the precise parameters of a particular feature", and that "faces in everyday environments are naturally obscured simply and extensively in various uncontrolled ways, by intervening objects, other people, shadows, the talker's own hand or hair" [4, p. 2271].

One such category of realistic facial occlusions is the various types of face-concealing garments and headgear worn for occupational, recreational, and religious purposes, or for the commission of crimes, such as assaults and robberies. It is these forms of occlusion that are of interest in the present study. The major difference from preceding research is that no post-production mask was applied to the video image (e.g. blacking out parts of the face), but that the talker's face was actually disguised while s/he was talking. As noted in [14], only very little research on AV speech processing when the talker's face is concealed by facewear has been carried out so far [15-17]. The advantage of this approach is that the auditory conditions in these studies reflect the articulatory and acoustic adjustments talkers might make to compensate for the obstruction, as well as the transmission loss caused by the mask material itself [14].

Aside from extending this line of research by including face concealments which are regularly found in real-world linguistic interaction, the present study aims to address the practical needs in casework carried out by forensic speech scientists. An appreciable proportion of forensic-phonetic casework involves facial disguise of one form or another (Peter French, York, personal communication). Hence, experts have to cope with speech samples produced through facewear on a fairly regular

basis. The current study is among the first to establish solid experimental data in which forensic speech experts can ground estimates of the influence such face concealments may have on the reliability of forensic-phonetic evidence produced in such cases [14, 17]. This sort of evidence may arise in form of lay earwitness testimony (where no speech recording is available), or acoustic recordings of speech produced through a face covering (which can form the basis of a professional auditory and acoustic analysis) [18, 19]. In either scenario the accuracy of the observations made by the (lay or expert) listener/viewer, and the magnitude of the changes involved in speech through facewear, need to be further ascertained.

To sum up, the goal of the two experiments introduced in the following sections was to determine how accurately lay listeners can identify consonants spoken through facewear during auditory-only and auditory-visual presentation of the speech stimuli (in quiet and noise), and to estimate how much – if any – visual speech information can still be extracted from the talker's face when crucial articulators were fully or partly disguised by a range of forensically-realistic face coverings.

2. Experiment 1

The first experiment tested the ability of phonetically-untrained listeners to identify syllable-onset consonants produced under different facewear conditions when presented in auditory-only (AO) or auditory-visual (AV) formats. It aimed to investigate the impact of various forms of facial occlusion on AO and AV consonant perception under (otherwise) optimal listening and viewing conditions. Experiment 1 thereby established a baseline which facilitates comparison with the results from a subsequent speech in noise experiment (Experiment 2; see Section 3 below).

2.1. Method

2.1.1. Test material

The test material employed in this study was extracted from the 'Audio-Visual Face Cover Corpus (AVFCC)' [14], which was recorded in a sound-treated TV studio at the Department of Theatre, Film and Television, University of York, UK. The talkers were seated in front of a plain green background, while two light sources were arranged to produce uniform illumination across their faces. They were asked to avoid marked head movements during the recordings, and not to wear spectacles or conspicuous jewellery so as to avoid possible reflection caused by the spotlights.

Of the three simultaneous continuous audio recordings made during each recording session, this study used those captured from a DPA 4066 Omnidirectional Headband Microphone, which was placed at approximately 2cm from the right-hand corner of each talker's mouth, and taped to the facewear with adhesive tape, if necessary. The audio streams were recorded with an Edirol R-4 Pro Portable 4 Channel Recorder and a Sound Devices 552 Portable Production Mixer, and saved in WAV format (48kHz, 768kbit/s, 16-bit signed integer PCM encoding). They were not normalised for amplitude, so as to maintain the level differences which naturally occur when speaking through any of the various types of facewear.

From the two simultaneous continuous HD colour video recordings (made with two Panasonic AG-HPX171E Camera Recorders), this study used the footage in which the talkers were facing the camera. The camera was positioned so that the images consisted of the talker's entire head and shoulders in the center of the screen. Given that the computer monitor for stimuli prompting was placed directly below the camera lens, the impression was given that the talkers were looking into the lens. The videos were cut into individual files containing one stimulus sentence each. Video (originally encoded using MEncoder, Xvid codec) and audio data were saved as AVI container files using Canopus Edius v5.51 (25f/s, 1280x720).

Two types of stimuli were produced from these recordings: auditory-only and auditory-visual, the former by extracting the audio stream from the videos using FFmpeg. The duration of all files was 2.2s. The high quality of the material allowed cues to fine phonetic detail in the talker's articulating face to be evident.

The speech material consisted of /C_1ɑ:C_2/ nonsense syllables embedded utterance-finally in the carrier sentence *He said* <*syllable*>. The sixteen consonants under investigation were /p b t d k g f v s z ʃ ʒ θ ð m n/. The target stimuli in this study were two tokens of each of these consonants in syllable onsets (/C_1/). To provide a consistent environment for consonant perception, the nucleus was always the open back vowel /ɑ:/ [8]. Logatoms were used so as to counter the effects of top-down processing (e.g. by virtue of lexical predictability) [6].

All ten talkers in the AVFCC corpus were included in this study. They were native English speakers who talked with a Southern Standard British English accent. Their average age was 27 (*SD* = 6). All of them had had previous training in the International Phonetic Alphabet, and none of them reported prior experience of wearing any type of facewear on a regular basis. The two tokens per consonant were produced by two different talkers to take into account idiosyncrasies and variability across talkers. To avoid bias, all subjects in the two perception experiments reported here were unfamiliar with the talkers.

All eight types of facewear included in the AVFCC corpus were tested in this study. These were: a balaclava with a mouth hole, a balaclava without a mouth hole, a motorcycle helmet, a hooded sweatshirt (hoodie) and scarf combination, a niqāb (full-face Muslim veil), a rubber mask, a surgical mask, and a piece of tape across the talker's mouth. The study also included a control condition (unconcealed face during the recordings) in order to provide a baseline for comparison with the results from the facewear conditions.

In sum, Experiment 1 tested consonant identification in two presentation modalities (auditory-only, auditory-visual). Within each modality there were nine facewear conditions (control + eight types of facewear). Each condition consisted of 32 items (16 consonants * 2 tokens), so that the entire test material was comprised of 576 test items.

2.1.2. Subjects

44 native English-speaking students (26 females, 18 males) were recruited at the University of York, UK. Their mean age was 20 (*SD* = 2). None of them reported a history of hearing impairment, and all had normal or corrected-to-normal vision. No subject reported previous experience of wearing any type of facewear, or interacting with people who do so, on a regular basis. All volunteers participated in the experiment in return for a small remuneration.

2.1.3. Procedure

The study was approved by the University of York Humanities and Social Sciences Ethics Committee. Prior to taking part, the

subjects were informed about the procedure so that they could grant their informed consent to participate. Both verbal and written instructions were given, and these were formulated in such a way as to avoid biasing the subjects towards one modality. Subjects were advised that the task in each trial of the forced-choice experiment was to identify only the initial consonant in the test syllable, and to then click one of the response items in a 2x8 grid presented on a computer screen. The response items showed the 16 consonants in orthographic representation (<p b t d k g f v s z sh zh th dh m n>), and additionally embedded in example words (minimal pairs where possible, i.e., *pit - bit*, *tie - die*, *kite - guide*, *few - view*, *sip - zip*, *she - genre*, *thin - this*, *map - nap*). The experiment was not timed; however, to help minimise the time taken to find the desired response, items were positioned in the grid according to their manner of articulation and voicing features. To familiarise the subjects with the experimental interface and procedure, they firstly completed a practice session (consisting of five AO and five AV control items), during which they also had the possibility of adjusting the playback volume to a comfortable hearing level. The main experiment was presented in three blocks, the presentation order of which was counterbalanced across subjects. Between each block subjects took a short rest break during which they had an informal conversation with the experimenter (first author). To compensate for practice and fatigue effects, the order of trials was pseudorandomised for each subject. No feedback about the correctness of responses was given to them. The experiment was run in a quiet computer lab at the Department of Language and Linguistic Science, University of York. Audio was played back through Sennheiser HD 280 PRO headphones, and videos were presented on a 22-inch Iiyama ProLite E2210HDS LCD monitor. The test was run using experimental control software specifically designed for the purpose of this study on the graphical framework wxLua. The entire experiment, including (de)briefing and breaks, took approximately 1.5 - 2hrs to complete.

2.2. Results

The performance measure calculated to express the subjects' ability to accurately identify the consonants was *percentage correct*. The accuracy scores were analysed by conducting a series of three-way repeated-measures analyses of variance (ANOVA) using IBM SPSS Statistics V.19.0.0.1, with modality (AO, AV), facewear (control, balaclava with and without mouth hole, hoodie/scarf, helmet, niqāb, rubber mask, surgical mask, tape), and consonant (/p b t d k g f v s z ʃ ʒ θ ð m n/) as independent within-group factors. All results were averaged across talkers. Where Mauchly's test indicated that the assumption of sphericity had been violated, the degrees of freedom and the *p*-values were adjusted using the Greenhouse-Geisser correction. Effects are reported as significant at $p < .05$. Lack of space rules out a discussion of the resulting patterns of consonant recognition errors, and whether these were consistent across the various experimental conditions. A detailed analysis of the consonant confusion matrices will, however, be presented in a future publication.

The data set produced by one female subject was excluded from the analysis as her results deviated significantly from the rest of the subjects (statistical outliers were defined as those falling into the 1.5 interquartile ranges below the 25th and above the 75th percentile). As the statistical analysis of the remaining data (24,768 observations in total) revealed, there was a weak but significant main effect of modality on consonant identification [$F(1,42) = 5.11$, $p < .05$, $\eta_p^2 = .11$], indicating that the subjects on average correctly identified more consonants when the talker's face could also be seen, as compared to when they only heard the talker's voice (see Figure 1). The main effects of facewear [$F(6,239) = 87.43$, $p < .001$, $\eta_p^2 = .68$] and consonant [$F(3,120) = 26.90$, $p < .001$, $\eta_p^2 = .39$] were also significant, as were the interactions between facewear and consonant [$F(120,5040) = 11.37$, $p < .001$, $\eta_p^2 = .21$], and between modality, facewear and consonant [$F(120,5040) = 1.23$, $p < .05$, $\eta_p^2 = .03$].

To explore the effects of facewear on the consonant ratings in further depth, ANOVAs were run for the AO and AV conditions, as well as for all facewear conditions separately. In the AO condition, the main effects of facewear [$F(6,245) = 56.20$, $p < .001$, $\eta_p^2 = .57$] and consonant [$F(3,134) = 27.65$, $p < .001$, $\eta_p^2 = .40$], and the facewear * consonant interaction [$F(120,5040) = 7.67$, $p < .001$, $\eta_p^2 = .15$], were significant. Similarly, in the AV condition, there was a significant main effect of facewear [$F(6,246) = 38.28$, $p < .001$, $\eta_p^2 = .48$] and of consonant [$F(3,119) = 23.78$, $p < .001$, $\eta_p^2 = .36$], as well as a significant facewear * consonant interaction [$F(120,5040) = 6.27$, $p < .001$, $\eta_p^2 = .13$]. In subsequent *post-hoc* Bonferroni-adjusted pairwise comparisons, the results pooled by facewear type were compared to the control condition, a test which sought to establish whether the subjects' performance in the various facewear conditions significantly differed from the baseline. It was found that in both the AO and AV conditions only the accuracy scores obtained for the tape condition significantly differed from the control (*p*s < .001). This indicates that AO and AV consonant identification accuracy significantly decreased when the speech was produced through the tape, but that none of the other face coverings significantly affected the subjects' performance.

Finally, the ANOVAs run for each facewear condition individually revealed a significant effect of modality again only for the tape [$F(1,42) = 6.45$, $p < .05$, $\eta_p^2 = .13$]. This implies that only when the speech was produced with the talker's mouth taped closed did speech intelligibility improve overall when visual speech cues were additionally available in the talker's articulating face (see the solid black versus black hatched bars in Figures 2-4).

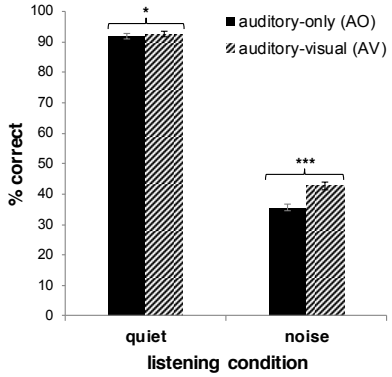


Figure 1: *Consonant identification accuracy averaged across facewear and consonants, for each listening condition separately (quiet = Experiment 1, noise = Experiment 2), as a function of modality. The error bars show the standard error of the mean. *** p < .001, * p < .05.*

3. Experiment 2

The second experiment built on the findings from Experiment 1 and presented the same set of stimuli, but in background noise. This again tested the subjects' ability to identify consonants spoken behind various face concealments when presented in AO or AV conditions, but this time the listening conditions were considerably degraded. The aim was to determine the contribution of facial speech cues when the subjects had to rely to a greater extent on the visual input owing to an expected decrease in auditory intelligibility.

3.1. Method

3.1.1. Test material

The same test material as that described for Experiment 1 was used in Experiment 2, with the exception that the audio streams in both the AO and AV conditions had background noise superimposed upon them. More specifically, 8-talker babble, which has been shown to reflect difficult listening conditions in a natural way, was used to mask the speech in this study [20]. The babble consisted of recordings of four females and four males speaking aloud while solving a Sudoku puzzle, from which pauses were removed, and which were normalised to the same RMS level before mixing. 30s of the resulting babble soundtrack was upsampled to 48kHz, and a random segment was selected to be added to each stimulus file. All noise fragments had the same RMS level when mixed with the speech. The original speech stimuli were 'on average' normalised. This means that at first the RMS energies of each talker's control samples were computed based on the *He said* frames of the test sentences. The mean RMS energy levels calculated from these multiple control samples per talker were then taken as the scale factors to normalise all speech samples (including the facewear conditions) on a per-talker basis. After that the rescaled speech was mixed with the babble noise using Matlab. The mixed files were not normalised, which means that the noise level was kept constant, and also that the natural variations in the speech levels (caused by the facewear) were again maintained during testing ($\bar{x}$ = -10.8dB SPL, *SD* = 4.8; calculated with pauses included). Finally, the visual test items were created by realigning the new 'noisy' audio streams with the original videos using VirtualDub 1.9.11.

3.1.2. Subjects

43 native English-speaking students (35 females, 8 males) from the University of Western Sydney, Australia, participated in the experiment. They were on average 20 years old (*SD* = 3) and reported normal or corrected-to-normal vision. None of them reported a history of hearing impairment, previous experience of regularly wearing any type of facewear, or interacting with people who do so. All subjects participated in return for course credit. The responses from two female and two male subjects had to be excluded from the test set owing to technical problems.

3.1.3. Procedure

The experiment was approved by the University of Western Sydney Human Research Ethics Committee. The procedure was the same as described for Experiment 1. Here, subjects were tested individually in a sound-attenuated IAC booth at the MARCS Institute, University of Western Sydney. Audio was played back through Sennheiser HD 650 headphones, and videos were presented on a 22-inch BenQ E2200HD LCD monitor.

3.2. Results

The data were analysed by means of three-way repeated-measures ANOVAs following the specifications given for Experiment 1. For the speech in noise data (22,464 observations in total) there were again significant main effects of modality [$F(1,38) = 196.12$, $p < .001$, $\eta_p^2 = .84$], facewear [$F(5,207) = 291.93$, $p < .001$, $\eta_p^2 = .89$] and consonant [$F(10,378) = 105.96$, $p < .001$, $\eta_p^2 = .74$] on the consonant ratings. The modality * facewear ($F(8,304) = 37.13$, $p < .001$, $\eta_p^2 = .49$), modality * consonant ($F(10,368) = 7.70$, $p < .001$, $\eta_p^2 = .17$), facewear * consonant ($F(120,4560) = 24.01$, $p < .001$, $\eta_p^2 = .39$), and modality * facewear * consonant ($F(120,4560) = 4.81$, $p < .001$, $\eta_p^2 = .11$) interactions were also all significant.

Once again, to explore the facewear effects further, ANOVAs were run for the AO and AV, and for the facewear conditions individually. In the AO condition, the main effects of facewear [$F(7,213) = 145.85$, $p < .001$, $\eta_p^2 = .79$] and of consonant [$F(15,570) = 80.30$, $p < .001$, $\eta_p^2 = .68$] were significant, as was the interaction between facewear and modality [$F(120,4560) = 14.65$, $p < .001$, $\eta_p^2 = .28$]. Likewise, in the AV condition, there was a significant main effect of facewear [$F(8,304) = 262.86$, $p < .001$, $\eta_p^2 = .87$] and of consonant [$F(10,370) = 94.68$, $p < .001$, $\eta_p^2 = .71$], and a significant facewear * modality interaction [$F(120,4560) = 18.06$, $p < .001$, $\eta_p^2 = .32$]. *Post-hoc* Bonferroni-adjusted pairwise comparisons revealed that in the AV condition the accuracy scores for all types of face coverings were significantly lower than in the control condition (*p*s < .001), indicating that the impoverished visual speech cues caused by all types of facial occlusions had a detrimental effect on consonant identification in noise. In the AO data, on the other hand, the recognition rates were significantly lower than the control only for certain types of facewear, namely the tape, the rubber mask, the helmet (*p*s < .001), the niqāb ($p < .01$), and the balaclava with the mouth hole ($p < .05$).

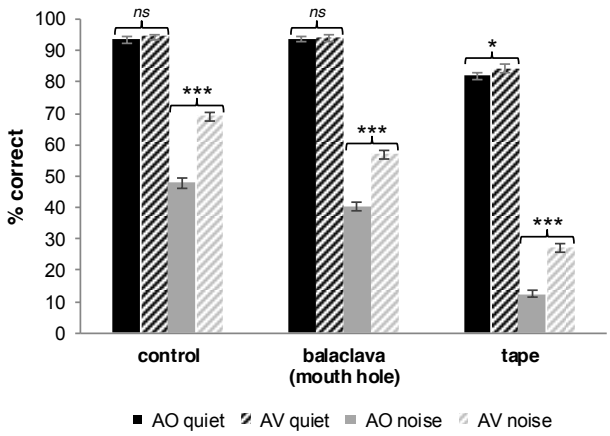


Figure 2: *Consonant identification accuracy averaged across consonants, for each listening condition, and for the control condition, the balaclava (with the mouth hole) and the tape separately, as a function of modality. The error bars show the standard error of the mean. *** p < .001, * p < .05, ns = non-significant.*

The ANOVAs run for each facewear condition separately were again aimed at determining whether having the talkers' faces visible led to an increase in recognition accuracy in each of the facewear conditions (including the control). This analysis revealed a significant main effect on consonant identification in the control condition [$F(1,38) = 146.09$, $p < .001$, $\eta_p^2 = .79$], the tape [$F(1,38) = 134.77$, $p < .001$, $\eta_p^2 = .78$], and the balaclava with the mouth hole [$F(1,38) = 130.64$, $p < .001$, $\eta_p^2 = .78$]. As can be seen in Figure 2 (solid grey and grey hatched bars), the gain in accuracy from the AO to the AV modality was significant, affirming a strong 'AV effect' in these conditions.

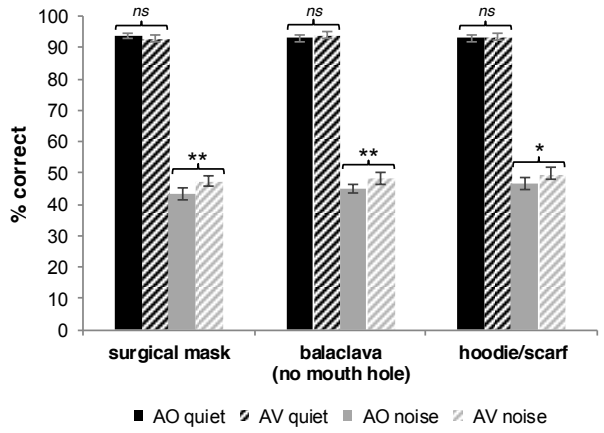


Figure 3: *Consonant identification accuracy averaged across consonants, for each listening condition, and for the surgical mask, balaclava (without the mouth hole), and the hoodie/scarf separately, as a function of modality. The error bars show the standard error of the mean. ** p < .01, * p < .05, ns = non-significant.*

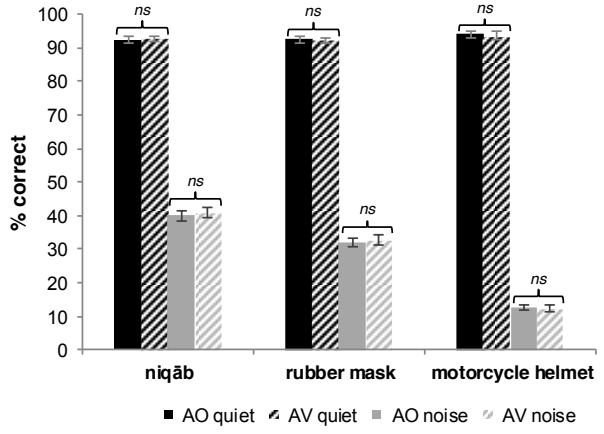


Figure 4: *Consonant identification accuracy averaged across consonants, for each listening condition, and for the niqāb, rubber mask, and the helmet separately, as a function of modality. The error bars show the standard error of the mean. ns = non-significant.*

There was a weaker but still significant difference between AO and AV consonant identification when the speech stimuli were produced through the balaclava without the mouth hole [$F(1,38) = 7.80$, $p < .01$, $\eta_p^2 = .17$], the surgical mask [$F(1,38) = 8.12$, $p < .01$, $\eta_p^2 = .18$], and the hoodie/scarf [$F(1,38) = 6.33$, $p < .05$, $\eta_p^2 = .14$]. As illustrated in Figure 3 (grey solid and grey hatched bars), in these conditions a significant rise in performance from AO to AV could be observed, but the AV effect was in each case less pronounced than for the control and the facewear shown in Figure 2.

Finally, no intelligibility gain when the videos of the talking heads were additionally presented to the subjects (i.e., no AV effect) was found for speech through the helmet ($p = .762$), the niqāb ($p = .488$), and the rubber mask ($p = .536$; see Figure 4).

4. General discussion

When the speech stimuli in the present study were presented in quiet listening conditions (Experiment 1), the subjects accurately identified the consonants in 92.2% of all cases, with hit rates ranging from 82.0% (tape/AO) to 94.4% (control/AV). By comparison, when the speech was embedded in 8-talker babble noise (Experiment 2), the hit rates markedly declined to overall 39.1% correct identifications (range: 12.4% for tape/AO, to 69.0% for control/AV).

In the quiet speech condition, a significant gain in consonant intelligibility was observed when visual speech information was presented simultaneously with the audio; however, the AV effect was overall weak. In fact, a detailed data analysis revealed that the effect only occurred in the tape condition, i.e., when the speech was produced and recorded while a piece of tape was adhered across the talker's mouth. For all other types of face and head coverings tested, the subjects' recognition accuracy did not significantly differ between the modalities; the subjects' performance was already near ceiling in the AO condition.

In the speech in noise experiment, on the other hand, the AO and AV hit rates varied as a function of facewear type quite substantially, and a significant AV effect was found only for a subset of the facewear. Moreover, the nine facewear conditions (including the control) evenly clustered into three categories. These differed with respect to the occurrence and strength of the AV effect, which in turn was directly related to the quantity of visual speech information recoverable from the talker's face.

The first category includes the control condition (absence of facewear), the balaclava with the mouth hole, and the tape across the talker's mouth (see Figure 2). The AV effect was strongest for these three conditions, which may for the most part be the result of lip motion still being somewhat visible to the observers. The possibility of lip-reading may have greatly aided consonant recognition, given that relevant phonetic cues – particularly to place of articulation – can be tracked from the talker's mouth region. As a side note, this was even possible for the tape, as the product used in this study was a 5cm wide, flexible surgical tape, which had been slightly loosened from the talker's lips so as to permit an airstream to escape from the talker's mouth.

The second category consisted of the surgical mask, the balaclava without the mouth hole, and the hoodie/scarf combination (see Figure 3). Statistical analysis still revealed a significant AV effect for these three types of facial occlusions, but the effect was less pronounced. This may have been the result of lip tracking no longer being possible, and possibly (also) of additional acoustic/auditory modifications to the speech signal (brought about by sound energy absorption caused by the mask material, and/or the facewear's interaction with the speech articulators [14, 15, 17]). However, AV consonant recognition

possibly increased in these cases as other forms of visual speech cues could still be extracted from the talker's moving face. For one thing, the comparatively tight fit of these types of facewear may have allowed the subjects to follow the talkers' jaw movements, which in turn may have drawn attention to critical events in the speech signal (such as identifying syllable onsets) [11, 21]. Previous research suggests that visual speech information can be widely distributed across the facial surface, and that the rapid contractions of muscles underlying the extraoral areas (e.g. chin, cheeks) are highly correlated with the movement of the oral articulators [4, 6-8, 13]. It has been shown that perceivers are able to extract even subtle phonemic features from a (human) face (unlike from e.g. an animated talking head, which typically does not encode these fine articulatory details [6]), and that in particular the chin and cheeks provide crucial information during speech processing (e.g. chin wrinkling; inflating of the cheeks at the sides of the mouth, near the upper lip, and at the side of the nose [6-8]). These types of visual cues may have provided helpful perceptual cues during the consonant identification exercise also for the subjects in the present study.

Finally, the third category of facewear included the niqāb, the rubber mask, and the motorcycle helmet (see Figure 4). Here, nearly the entire face was covered, with the result that these types of facewear neither allowed for lip-reading, nor for the extraction of any speech movements from the articulating face. Hence, it is perhaps unsurprising that no AV effect could be observed for the types of face concealment.

5. Conclusions

The present study established consonant recognition accuracy scores obtained from phonetically-untrained observers who participated in an auditory-only and auditory-visual consonant recognition experiment where the talker's face was obscured by one of eight types of forensically-realistic face coverings. It was found that facewear which allowed the viewer to recover lip movements promoted consonant perception accuracy, and that when the movements of the mouth were obscured, accuracy was still marginally improved by the provision of a visual image of the talker. Perceivers therefore appear to have made effective use of extraoral facial cues to consonant identity. This work extends previous research on auditory-only and auditory-visual speech perception in quiet and noise, and offers new insights into the effects of realistic facial occlusions on consonant recognition. By contrast with preceding research, which mainly tested the relevance of precisely defined facial areas during AV speech processing, the current study enhanced the naturalness of the AV speech material by testing a large variety of face and head coverings which are routinely, and in comparatively uncontrolled ways, encountered in real-life communicative situations.

6. Acknowledgements

This research has received funding from the European Commission's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 238803 (Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics [BBfor2]'). The first author would like to thank Martin Cooke for providing the babble noise stimuli, and for his very valuable advice. Thanks also to Chris Davis, Benjamin Schultz, Michael Fitzpatrick, Jeesun Kim, David van Leeuwen and Peter French for help and feedback, and to all those who talked, watched, and listened.

7. References

[1] Cotton, J. C., "Normal 'visual hearing'," *Science*, vol. 82, pp. 592-593, 1935.

[2] Greenberg, H. J. & Bode, D. L., "Visual Discrimination of Consonants," *J Speech Lang Hear Res*, vol. 11, no. 4, pp. 869-874, 1968.

[3] Davis, C. & Kim, J., "Audio-visual speech perception off the top of the head," *Cognition*, vol. 100, no. 3, pp. B21-B31, 2006.

[4] Jordan, T. R. & Thomas, S. M., "When half a face is as good as a whole: effects of simple substantial occlusion on visual and audiovisual speech perception," *Atten Percept Psychophys*, vol. 73, no. 7, pp. 2270-85, 2011.

[5] Kim, J., Davis, C. & Groot, C., "Speech identification in noise: Contribution of temporal, spectral, and visual speech cues," *J Acoust Soc Am*, vol. 126, no. 6, pp. 3246-57, 2009.

[6] Lidestam, B. & Beskow, J., "Visual Phonemic Ambiguity and Speechreading," *J Speech Lang Hear Res*, vol. 49, no. 4, pp. 835-847, 2006.

[7] Marassa, L. K. & Lansing, C. R., "Visual Word Recognition in Two Facial Motion Conditions: Full-Face Versus Lips-Plus-Mandible," *J Speech Lang Hear Res*, vol. 38, no. 6, pp. 1387-94, 1995.

[8] Preminger, J. E., Lin, H.-B., Payen, M. & Levitt, H., "Selective visual masking in speechreading," *J Speech Lang Hear Res*, vol. 41, no. 3, pp. 564-575, 1998.

[9] Rosenblum, L. D., "Primacy of Multimodal Speech Perception," in *The Handbook of Speech Perception*, Pisoni, D. B. & Remez, R. E., Eds., Oxford: Wiley-Blackwell, 2005, pp. 51-78.

[10] Rosenblum, L. D. & Saldaña, H. M., "An audiovisual test of kinematic primitives for visual speech perception," *J Exp Psychol Hum Percept Perform*, vol. 22, no. 2, pp. 318-331, 1996.

[11] Schwartz, J.-L., Berthommier, F. & Savariaux, C., "Seeing to Hear Better: Evidence for Early Audio-visual Interactions in Speech Identification," *Cognition*, vol. 93, no. 2, pp. B69-B78, 2004.

[12] Swerts, M. & Krahmer, E., "Facial expression and prosodic prominence: Effects of modality and facial area," *Journal of Phonetics*, vol. 36, no. 2, pp. 219-238, 2008.

[13] Thomas, S. M. & Jordan, T. R., "Contributions of oral and extraoral facial movement to visual and audiovisual speech perception," *J Exp Psychol Hum Percept Perform*, vol. 30, no. 5, pp. 873-888, 2004.

[14] Fecher, N., "The 'Audio-Visual Face Cover Corpus': Investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear," in *Proc of InterSpeech*, Portland, Oregon, USA, 2012.

[15] Fecher, N. & Watt, D., "Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives," in *Proc of ICPhS*, Hong Kong, China, 2011, pp. 663-666.

[16] Heath, A. J. & Moore, K., "Earwitness Memory: Effects of Facial Concealment on the Face Overshadowing Effect," *Int Journal of Advanced Sci and Technology*, vol. 33, pp. 131-140, 2011.

[17] Llamas, C., Harrison, P., Donnelly, D. & Watt, D., "Effects of different types of face coverings on speech acoustics and intelligibility," *York Papers in Linguistics (Series 2)*, vol. 9, pp. 80-104, 2008.

[18] Yarmey, A. D., "Factors Affecting Lay Persons' Identification of Speakers," in *The Oxford Handbook of Language and Law*, Tiersma, P. & Solan, L., Eds., Oxford, New York: Oxford University Press, pp. 547-556, 2012.

[19] Foulkes, P. & French, P., "Forensic Speaker Comparison," in *The Oxford Handbook of Language and Law*, Tiersma, P. & Solan, L., Eds., Oxford, New York: Oxford University Press, pp. 557-572, 2012.

[20] Simpson, S. A. & Cooke, M., "Consonant identification in N-talker babble is a nonmonotonic function of N," *J Acoust Soc Am*, vol. 118, no. 3, pp. 2775-78, 2005.

[21] Nahorna, O., Berthommier, F. & Schwartz, J.-L., "Binding and unbinding the auditory and visual streams in the McGurk effect," *J Acoust Soc Am*, vol. 132, no. 2, pp. 1061-1077, 2012.