Distorted visual information influences audiovisual perception of voicing

Ragnhild Eg, Dawn Behne

Department of Psychology, Norwegian University of Science and Technology

eg@stud.ntnu.no, dawn.behne@svt.ntnu.no

Abstract

Research has shown that visual information becomes less reliable when images are severely distorted. Furthermore, while voicing is generally identified from acoustical cues, it may also provide perception with visual cues. The current study investigated the impact of video distortion on the audiovisual perception of voicing. Audiovisual stimuli were presented to 30 participants with the original video quality, or with reduced video resolution (75x60 pixels, 45x36 pixels). Results revealed that in addition to increased auditory reliance with video distortion, particularly for voiceless stimuli, perception of voiceless stimuli was more influenced by the visual modality than voiced stimuli.

Index Terms: audiovisual perception, speech, voicing, visual distortion

1. Introduction

Previous research on audiovisual (AV) perception has repeatedly shown that speech comprehension becomes increasingly difficult with higher levels of auditory noise and that visual information carries more influence when the auditory modality is masked by noise (e.g., [1], [2], [3], [4]). Visual disturbances also affect AV perception, and speech comprehension suffers when available visual information is limited [5]. When the visual modality becomes too uninformative, perception tends to rely solely on auditory information [2]. The current study considers how video distortion affects AV perception of voicing and how the influence from the visual modality is reduced as the degree of distortion increase. The main focus is the effect of poor video resolution quality on the perception of voicing.

Numerous methods have been applied to create visual disturbances, for instance, affixing sheets of paper to a computer monitor [2] or manipulating the facial orientation [6]. However, a more common method is to manipulate video quality. Filters can be used to change the spatial frequency of the video imagery [7], while spatial degradation can be achieved through mosaic transformation ([8], [9]) or by manipulating pixel resolution (e.g., [10], [11]). Low pixel resolution is a particularly relevant issue for the continuing development of video conferencing, where compromises are still made between video quality and speed of transmission [12]. Although results typically show that perceptual reliance on vision is reduced with distortion, a common finding is that perception does not require high resolution images to make use of visual speech information (e.g., [7], [9]). While these methods all succeed in decreasing the reliability of the visual modality, they do not take into account how visual disturbances occur naturally in everyday life. In the current study, video quality is reduced by manipulating pixel resolution in order to blur images in a way that simulates the natural experience of poor eyesight or seeing at long distances.

While many have investigated the impact of visual disturbances on speech comprehension (e.g., [7], [13]) and the perception of place of articulation (POA) (e.g., [2], [6]), earlier studies have not specifically focused on how reduced visual information affects the perception of voicing. Voiced and voiceless stop consonants (stops) are distinguishable through differences in voice onset time (VOT), the interval between release burst and onset of vocal cord vibration [14]. Voiceless stops have longer VOTs and are considered more salient than voiced stops due to this longer delay prior to voicing [15]. However, voicing may not be solely acoustical in nature; the differences in the duration of VOT between voiced and voiceless consonants may provide perception with visual cues about voicing. Furthermore, the perception of voicing is closely linked to the perception of POA [16]; as visual cues contribute greatly to the identification of POA [17], they may also assist in distinguishing voiced and voiceless stops. If the longer VOTs of voiceless stops make them acoustically more salient and these do indeed provide perception with visual cues, they may well be more visually salient than voiced stops. Moreover, voiceless contrasts, with audio and video of incongruent POA, have been found to elicit more combination responses than corresponding voiced contrasts (e.g., [17], [18]). These results are consistent with voiceless stops being visually and acoustically more salient. For combination responses both the auditory and the visual modality are perceived, as opposed to fusion responses where audio and video are integrated into a single intermediary speech sound. The latter is more common for voiced stimuli (e.g., [17], [18]). Thus, whereas voiced stops are more susceptible to perceptual fusion, both auditory and visual information appear to influence the perception of voiceless stops.

In the current study, the differential use of auditory and visual information in AV perception of voicing is assessed using incongruent AV stimuli with differing degrees of visual distortion. As the visual saliency of voiceless stops is expected to be greater than for voiced stops, voiceless stimuli are expected to be more influenced by visual information than voiced stimuli. However, as voiceless stops have also been found to be acoustically more salient than voiced stops [15], voiceless stimuli are expected to be more affected by video distortion than voiced stimuli and thus show a greater increase in the use of auditory information. Moreover, video distortion should lead to an increase in responses influenced by the auditory modality, irrespective of voicing.

2. Method

Two sets of incongruent AV stimuli were used to assess the respective influences of the auditory and the visual modalities. For one set, audio and video components of the stimuli differed in voicing (AV(voiceing) stimuli); for the other set, audio and video differed in POA (AV(POA) stimuli). Moreover, congruent stimuli (AVc), as well as video only (V only) and audio only (A only) stimuli, served as controls for the validity of results and the intelligibility of stimuli.

2.1. Participants

Thirty participants (15 males and 15 females) aged between 19 and 32 years (M=22, SD=3.64), were recruited at the
Norwegian University of Science and Technology. All participants had Norwegian as their native language and reported normal hearing and normal-to-corrected vision.

2.2. Stimuli

Stimuli were created from audio and video recordings of a male speaker producing six consonant-vowel (CV) syllables (/ba/, /da/, /ga/, /pa/, /ta/, /ka/). Video distortion was achieved by manipulating the original resolution of 720x576 pixels and reducing it to either 75x60 pixels or 45x36 pixels. To avoid graininess, the resolution was then set back to the original level. The still shots in Figure 1 are examples of stimuli with and without manipulations. Resolution was adjusted for all AVvoicing, AViPOA, AVc and Vonly stimuli.

Figure 1. Illustrations of the video stimuli at (a) the original resolution (b) 75x60 pixels and (c) 45x36 pixels.

2.2.1. AVc, Vonly and Aonly stimuli

For AVc stimuli, audio and video recordings for corresponding CVs were left as they were, while video recordings were presented without audio for Vonly stimuli and audio recordings were presented with a black screen for Aonly stimuli. This resulted in six AVc, six Vonly and six Aonly stimuli, one for each CV. Aonly stimuli were less frequent than the remaining stimuli, as they were not presented at different levels of resolution.

2.2.2. AVvoicing stimuli

Audio and video recordings of all six CVs were combined and counterbalanced so that stimuli matched in POA, but differed in voicing. Thus six stimuli pairings were created (/ba-pa/, /pa-ba/, /da-ta/, /ta-da/, /ga-ka/, /ka-ga/).

2.2.3. AViPOA stimuli

Audio and video recordings of the four labial and velar CVs were combined and counterbalanced so that stimuli matched in voicing, but differed in POA. Thus four stimuli pairings were created (/ba-ga/, /ga-ba/, /pa-ka/, /ka-pa/).

2.3. Procedure

Participants were each seated approximately 50 cm in front of a 17-inch iMac monitor where visual components of stimuli were presented. Audio was presented over AKG K271 studio headphones. All stimuli were repeated twice; they were presented in two blocks and randomised across stimulus type within each block. For each trial the trial number was first displayed on the monitor, followed by the stimulus segment, after which participants indicated which of the six CVs (/ba, da, ga, pa, ta, ka) was perceived. Participants pressed the ‘enter’ button on the keyboard to proceed to the next trial.

3. Results and Discussion

3.1. AVc, Vonly and Aonly stimuli

Responses to AVc, Vonly and Aonly stimuli were each tabulated based on whether they matched the presented stimuli. Percent correct was thereby calculated as a measure of accuracy and used as a dependent variable for the separate repeated measures ANOVAs that were run for AVc and Vonly stimuli with voicing and resolution as independent variables, as well as for a paired-samples t-test run for Aonly stimuli with voicing as the single factor. Here and in later analyses, significant main effects were followed up using paired samples t-tests, while significant interactions were further analysed using Tukey’s Honestly Significant Difference test (Tukey’s T-test).

For both AVC and Aonly stimuli responses were close to 100% correct across conditions, whereas response accuracy for Vonly stimuli varied according to both voicing and resolution. For AVc stimuli no reliable differences were observed for voicing [$F(1,29)=0.33$, ns], resolution [$F(2,58)=0.24$, ns], or their interaction [$F(2,58)=1.17$, ns]. Nor for Aonly stimuli was there a significant difference in voicing [$t(29)=1.00$, ns]. However, reliable effects of voicing [$F(1,29)=5.64$, $p=.024$] and resolution [$F(2,58)=10.63$, $p=.001$] were observed for Vonly stimuli, as shown in Figure 2, although the interaction between voicing and resolution was non-significant [$F(2,58)=1.81$, ns]. The main effects revealed that response accuracy was greater for voiced than for voiceless stimuli, while for resolution, accuracy was better at the original resolution compared to 75x60 pixel resolution [$t(29)=3.75$, $p=.001$] and 45x36 pixel resolution [$t(29)=4.29$, $p=.001$]. The difference between 75x60 pixels and 45x36 pixels was not significant [$t(29)=0.52$, ns].

While video distortion had no effect on congruent stimuli, the reduced response accuracy for Vonly stimuli between the original resolution and 75x60 pixels indicates that the video manipulations had an impact on visual perception and that reducing pixel resolution is a valid method for creating visual distortion. Moreover, with approximately 50% correct for Vonly stimuli at the original resolution, participants are responding with far greater accuracy than mere chance (16.67%), indicating that video recordings are intelligible and of sufficient quality. The near-perfect accuracies for Aonly stimuli suggest that audio recordings are clear and also have adequate signal quality. Furthermore, accuracy scores nearing 100% for congruent stimuli, indicate that the AV presentations of stimuli are non-confounded and intelligible.

3.2. AVvoicing stimuli

Responses to AVvoicing stimuli were coded as matching with the auditory modality (audio match) or matching with the visual modality (video match) of the presented stimuli. Separate repeated measures ANOVAs were run for percent
audio match and percent video match with voicing and resolution as independent variables.

For AVvoicing stimuli, audio match was close to 100% across conditions, with no reliable effects of voicing \([F(1,29)=3.92, \text{ ns}]\), resolution \([F(2,58)=3.01, \text{ ns}]\) or interaction \([F(2,58)=1.91, \text{ ns}]\). Correspondingly, video match scores were close to 0% across conditions, and here again no reliable differences were observed for voicing \([F(1,29)=0.00, \text{ ns}]\), resolution \([F(2,58)=2.07, \text{ ns}]\), or their interaction \([F(2,58)=0.00, \text{ ns}]\).

This shows that responses to AVvoicing stimuli are affected by the auditory modality, with no effect of voicing or resolution. While visual cues were expected to differ for voiced and voiceless stops, voicing was found to be mainly perceived from acoustical cues [14]. The lack of influence from the visual modality clearly indicates that perception favours auditory information when presented with conflicting AV voicing signals, even when the audio is voiced.

### 3.3. AViPOA stimuli

For AViPOA stimuli, percent audio match and video match were calculated, as well as the percent of “fusion” responses which were intermediate the POA of the audio and video components of the stimuli and matched the voicing of the presented stimuli. Separate repeated measures ANOVAs were carried out for audio match, video match and fusion scores with voicing and resolution as independent variables.

#### 3.3.1. Voicing

With AViPOA stimuli, main effects for voicing were revealed for audio match \([F(1,29)=22.59, p=0.001]\), video match \([F(1,29)=6.43, p=0.017]\) and fusion responses \([F(1,29)=6.75, p=0.015]\). Means for all response types for voiced and voiceless stimuli are presented in Figure 3.

![Figure 3. Mean percentages for audio match, fusion and video match responses for voiced and voiceless AViPOA stimuli.](image)

Voiceless

Compared to voiced stimuli, voiceless stimuli had a greater share of video match and fusion responses, and a correspondingly smaller percentage of audio match responses. As expected, results indicate that voiceless stimuli are more influenced by visual information than voiced stimuli, although the perception of both is mainly cued by the auditory modality. Voiceless consonants have been argued to be more salient than voiced consonants due to longer VOTs [15]. While this saliency is based on findings for acoustic attributes [14], it may also be visually realised in the speakers face. If VOT contributes with visual information, then voiceless stops should also be more visually salient. Furthermore, the perception of POA is closely linked to voicing [16]. The visual cues that help identify POA may vary between voiced and voiceless stops, providing visually perceptible cues to voicing.

#### 3.3.2. Resolution

Reliable effects of resolution were also observed for audio match \([F(2,58)=24.39, p=0.001]\), video match \([F(2,58)=6.86, p=0.002]\) and fusion responses \([F(2,58)=11.74, p=0.001]\). Means for all response types, across resolution levels, are presented in Figure 4. Audio match scores were less frequent at the original resolution than at 75x60 pixels \([n(29)=3.60, p=0.001]\) and at 45x36 pixels \([n(29)=6.00, p=0.001]\), as well as at 75x60 pixels compared to 45x36 pixels \([n(29)=4.26, p=0.001]\). Video match scores were more frequent at the original resolution compared to 75x60 pixels \([n(29)=2.54, p=0.017]\) and 45x36 pixels \([n(29)=2.85, p=0.008]\), although the difference between 75x60 pixels and 45x36 was not significant \([n(29)=1.98, \text{ ns}]\). Fusion responses were more frequent at the original resolution than at 75x60 pixels \([n(29)=2.36, p=0.025]\) and at 45x36 pixels \([n(29)=4.25, p=0.001]\), and were also more frequent at 75x60 pixels compared to 45x36 pixels \([n(29)=3.00, p=0.005]\).

These findings suggest that video distortion increases the influence from the auditory modality, while reliance on the visual modality decreases, in accordance with expectations and previous research [2], [13]). Furthermore, even when video was presented with a resolution of 45x36 pixels, a share of responses was still influenced by the visual modality. With only vision to rely on, over a third of responses were accurate for V-only stimuli at this level of resolution. This tendency may be the outcome of perception not requiring all the information that is available in high resolution images [7]. Since fine-grained details are not essential, it could be that speech quality is provided by the overall unity of visual details [9]. Although relevant information may be lost when video quality is reduced, perception may be able to compensate for this loss of detail as long as the outlines of facial movements remain perceptually accessible. An implication of this is that the compromise between video quality and speed of transmission for video conferencing systems should favour speech and synchrony between audio and video.

#### 3.3.3. Voicing and resolution

The interaction between voicing and resolution was significant for audio match responses \([F(2,58)=3.84, p=0.027]\), but not for video match \([F(2,58)=0.86, \text{ ns}]\) or fusion responses \([F(2,58)=2.12, \text{ ns}]\). Means for voiced and voiceless audio match responses, across resolution levels, are presented in Figure 5. Results from Tukey’s T-test \([T(1,29)=8.88, p<0.05]\) revealed that audio match responses were more frequent for

![Figure 4. Mean percentages for audio match, fusion and video match responses for AViPOA stimuli at the original resolution, 75x60 pixels and 45x36 pixels.](image)
Voiced than for voiceless stimuli at the original resolution. Moreover, voiceless stimuli had a greater share of audio match responses at 45x36 pixels compared to 75x60 pixels and the original resolution, and at 75x60 pixels compared to the original resolution. Whereas for voiced stimuli, the difference was only significant between 45x36 pixels and the original resolution.

![Graph](image)

Figure 5. Mean percentages for audio match responses for voiced and voiceless AV/POA stimuli at the original resolution, 75x60 pixels and 45x36 pixels.

Voiceless stimuli were expected to be more affected by video distortion and show a greater increase in the use of auditory information, compared to voiced stimuli. This expectation is supported by the increased use of auditory cues for voiceless stimuli as resolution quality decreased. For voiced stimuli, the increase in influence from the auditory modality was only significant between the finest and poorest resolution quality. Furthermore, voiceless stimuli were found to be cued less by the auditory modality than voiced stimuli at the original resolution, but not at 75x60 pixels and 45x36 pixels. This could be an indication that the longer VOTs for voiceless stops [15] are also visually accessible and that the loss of visual information with video degradation is greater for voiceless syllables than for voiced syllables. Although voiceless stops may be more visually salient than voiced stops, the high rate of responses matching the auditory modality indicates that voicing is nevertheless very much acoustic in nature.

4. Conclusions

This study has focused on how degraded visual information is used in perception and how the influence from the auditory modality is exaggerated when the visual modality becomes unreliable. Furthermore, it has looked at the degree to which perception can make use of visual cues about voicing and whether voiced and voiceless speech sounds differ in available visual speech information. Of special interest was the finding that voiceless consonants were more influenced by visual information than voiced consonants. This suggests that visual cues must be available through voiceless speech sounds and can provide perceptual cues about the voicing state of a consonant. These visual cues may be related to the duration of VOT [15], or they may be linked to the production of POA [16]. Voicing is nevertheless more of an acoustic characteristic of speech than a visual one [14]. When video images were presented at lower resolutions, the visual modality had a reduced perceptual influence. This reduction was more dramatic for voiceless stimuli that depended more on visual information than voiced stimuli. With a resolution at 45x36 pixels, nearly all responses reflected the auditory modality. Thus perception may use visual speech information to process voicing, particularly voiceless sounds, but as soon as the visual signal becomes distorted, the auditory signal will override it. Moreover, this study did not show a visual effect on perception when incongruent stimuli differed in voicing. No matter which modality was voiced and which was voiceless; the auditory signal generally provided the predominant cue. Whatever visual cues may be available for voicing, acoustic cues remain the more useful to perception.

5. References


