Voicing influences the saliency of place of articulation in audio-visual speech perception in babble

Magnus Alm, Dawn Behne

Department of Psychology, Norwegian University of Science and Technology

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

Abstract

Previous research has shown that voicing can influence the perception of consonant place of articulation (POA) in audio-visual (AV) speech perception, although findings are inconsistent and often differ with the use of background noise. The prediction in the current study was that the AV perception of voiced and voiceless stop consonant POA is influenced by the differences in spectral distribution between voiced and voiceless stops, a hypothesis not compatible with the direction of the voicing effect shifting with different types of noise. Fifteen young adults were tested using incongruent AV stimuli that differed in POA, in a voiced and a voiceless condition, applying the infrequently used babble noise as background. As predicted participants used the auditory modality to a greater extent identifying the POA of voiced stops compared to voiceless stops. The more distinct spectral distribution of the voiced stops may contribute to them being more easily identified auditorily on the POA dimension than the voiceless stops. The study extends previous research using white noise, by demonstrating a consistent pattern of results in babble noise.

Index Terms: Speech perception, place of articulation, voicing, babble.

1. Introduction

Research shows that in audio-visual (AV) perception of consonant place of articulation (POA) and voicing, audio and visual modalities are used to different extents. While POA is very susceptible to auditory noise it is visually salient, which differs from consonant voicing which is auditorily more robust but not visually salient [e.g., 1, 2, 3]. A study of audio-visual perception is thus more suited to assess susceptibility to auditory noise than an audio-only design because it offers a complementary route to perception and thus better reflects auditory uncertainty by a perceiver.

Previous research has indicated that perceived POA and voicing of a consonant influence each other [e.g., 4, 5]. For example, voice onset time (VOT), considered the most important cue for voicing [6], has been shown to influence the perception of POA [7]. Research has, however, typically isolated one of the two dimensions for analysis [e.g., 8, 9, 10]. While the interdependency between voicing and POA has seldom been explicitly considered, it is inevitable for perceivers, for whom the aggregate of the fragmental articulatory building blocks ultimately give rise to a meaningful holistic percept (e.g., consonant phonemes).

The formant transition between a consonant and a vowel in a syllable is an important cue for identification of POA [11]. Research shows that, at voice onset, the formant frequencies differ for oral stops differing in POA [12] and this distinctiveness may aid in discriminating, for example, a labial /b/ from a velar /g/. This auditory distinctiveness is however, less prominent for voiceless stops than it is for voiced stops. In particular, at the onset of voicing, $F_2$ and $F_1$ of the voiceless labial /p/ are more spectrally similar to the voiceless velar /k/ than the voiced labial /b/ is to the voiced velar /g/ [12]. The few studies that explicitly consider voicing’s influence on the perception of POA do however yield somewhat conflicting results. Whereas some studies indicate that POA is more auditorily salient for voiced consonants [14], others indicate that POA is more auditorily salient for voiceless consonants [3, 13].

This reported discrepancy may be attributed to different use of noise; that is, no noise [e.g., 13], babble noise [e.g., 3] and white noise [e.g., 14]. Experimental assessment of the noise susceptibility of POA and voicing is typically conducted by varying the levels (dB) of white noise [e.g., 2, 8, 14], while the influence of different types of noise is seldom considered. Audio-only studies show that babble noise possesses characteristics that distinguish it from white noise [e.g., 15, 16]. Differences may be related to spectral distribution [16, 17] and semantic interference [15].

Previous research has seldom explicitly addressed the interdependency of voicing and POA in perception of stop consonants in noise, and when this interdependency is considered they often show inconsistent results. The current study hence investigates the influence of voicing on perceived consonant POA in babble noise. Babble noise is applied because although it is a common and naturally occurring distractor in speech perception it is infrequently used in AV research, and because it enables a direct comparison of results with that of Behne et al. (2006). POA is expected to be more auditorily salient for voiced stops than for voiceless stops in noise due to the voiced stops being more spectrally distinct than the voiceless stops [12]. Voiced stops are predicted to yield more auditory responses than the more noise susceptible voiceless stops. Voiceless stops’ susceptibility to noise is expected to provoke a shift towards a more reliable route of perception in noise (i.e., the visual modality) and thus be reflected in more visual responses.

2. Method

Fifteen young adults within the range of 20 to 31 years of age ($M=23, SD=7$) participated in the experiment. The group consisted of 7 males and 8 females. All participants were Norwegian native speakers and reported normal hearing and normal or corrected to normal vision. All participants also had an eastern dialect of Norwegian.

Stimuli were developed from AV recordings of a male speaker with an urban Eastern-Norwegian dialect. The AV recordings consisted of the consonant-vowel syllables /ba, ga, pa/ and /ka/. The incongruent stimuli were made by dubbing the velar visual syllables onto the labial auditory
syllables with the same consonant voicing (i.e., \(A_vV_a\) and \(A_aV_v\)). Hence an AV stimulus was either voiced or voiceless and the A and V components differed only in the POA dimension. The syllables were presented in quiet and in babble noise of -12 dB SNR with the same duration as the syllable.

Stimuli were visually presented on a 17” monitor (1440x900 pixels) and sound was conveyed to a set of AKG K271 stereo closed dynamic circumaural studio headphones, at a 68 ± 2 dB level (corresponding to a frontally incident free-field sound of 68 dBA).

The stimuli were internally randomized in four independent blocks. Each stimulus was presented four times, but appeared only once in a block. Stimuli were presented in a forced choice task where, for each incongruent syllable, participants indicated which of six alternatives (ba, ga, pa, ka, da and ta) best corresponded to what was perceived.

### 3. Results

Participants POA judgements were tabulated according to whether the consonant matched the audio velar component (A-match), the visual labial component (V-match) or the alveolar intermediate to the audio and visual components (AV-fusion).

Data were analyzed with a repeated measures analysis of variance (ANOVA), where voicing (voiced, voiceless) and noise (quiet, babble) were independent variables and A-match, V-match and AV-fusion were dependent variables. The ANOVA results are reported in Table 1 and illustrated in Figure 1.

#### Table 1: Statistical results of the ANOVA showing main effects and the interaction for voicing and noise.

<table>
<thead>
<tr>
<th>Factor(s)</th>
<th>df</th>
<th>F</th>
<th>p</th>
<th>F</th>
<th>p</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voicing</td>
<td>1,14</td>
<td>42.1</td>
<td>&lt;.001</td>
<td>33.8</td>
<td>&lt;.001</td>
<td>1.0</td>
<td>n.s</td>
</tr>
<tr>
<td>Noise</td>
<td>1,14</td>
<td>168.0</td>
<td>&lt;.001</td>
<td>217.6</td>
<td>&lt;.001</td>
<td>1.0</td>
<td>n.s</td>
</tr>
<tr>
<td>Voicing*Noise</td>
<td>1,14</td>
<td>62.5</td>
<td>&lt;.001</td>
<td>51.9</td>
<td>&lt;.001</td>
<td>1.0</td>
<td>n.s</td>
</tr>
</tbody>
</table>

Figure 1: Percent A-match, V-match and AV-fusion for voiced and voiceless stimuli in quiet and babble noise.

Figure 1 shows that participants POA judgements used the auditory component significantly more, and almost exclusively, in the quiet condition (\(M= 99 \%\), \(SE= 0.83\)) compared to the babble noise condition (\(M=49\%\), \(SE= 4.47\)). The POA responses also match the auditory component to a greater extent for voiced consonants (\(M= 93\%\), \(SE=4.2\)), than for voiceless consonants (\(M=55\%\), \(SE= 3.62\)). Fisher’s Protected Least Significant Difference tests (LSD) \((p<.05)\) revealed that in babble noise voiced consonants resulted in significantly more auditory responses than voiceless consonants \((p<.001)\), whereas no difference between voicing conditions was obtained in quiet. Babble noise resulted in less auditory responses than the quiet condition for voiceless stimuli \((p<.001)\), but not for voiced stimuli.

POA responses match the visual component to a significantly lesser extent in the quiet condition (\(M= 1\%\), \(SE= 0.83\)) than in the babble noise condition (\(M= 50\%\), \(SE= 3.86\)) and participants gave fewer visual responses for voiced consonants (\(M= 7\%\), \(SE= 4.2\)) than for voiceless consonants (\(M=44\%\), \(SE=3.63\)). LSD tests showed significantly more visual responses for voiceless stimuli than for voiced stimuli only in babble noise \((p<.001)\). Babble noise attributed to more visual responses than the quiet condition for voiceless consonants \((p<.001)\), but not for voiced consonants.

For AV-fusion no significant differences in POA responses were found for voicing or noise.

### 4. Discussion

In quiet, POA responses almost solely matched the auditory component, whereas in babble, visual input supplements speech perception, a finding consistent with previous research investigating AV perception of incongruent \(A_{velar}V_{labial}\) stimulis in quiet conditions [e.g., 1, 2, 3]. In babble noise however, the results clearly show an influence of voicing on the use of modality in POA perception. Here perceivers use the auditory modality to a significantly greater extent for voiced stimuli than for voiceless stimuli. Although consistent with the study’s predictions, it contrasts some previous research [3, 13].

The current study is based on that of Behne et al. (2006) which used the same speaker and noise recordings. Surprisingly, in the Behne et al. (2006) study participants used the auditory modality more for voiceless stimuli than voiced stimuli when perceiving POA. The target stimuli of the two experiments differed in one important respect: whereas the current experiment used only one vowel context (\(/a/\)), Behne et al. 2006 used two (/a/ and /i/). When analyzing the influence of voicing on perception of POA, Behne et al. (2006) collapsed data across vowel contexts. Several studies [e.g., 18, 19] show that the vowel context influences the perception of consonants. Analyzing the data from the Behne et al. (2006) study in regard to vowel context revealed results consistent with that of the current study: in the /a/ vowel context participants indeed used the auditory modality to a greater extent for voiced stimuli than for voiceless stimuli.

Taken together, these studies indicate that voiced consonants facilitate the auditory perception of POA in /a/-vowel contexts. How can this phenomenon be explained? The results show a clear-cut effect of noise, with the quiet condition yielding no differences between voiced and voiceless consonants. Similar results obtained by Grauwinkel using white noise indicate that this phenomenon is not exclusively associated with babble noise [14]. The voicing effect in the Grauwinkel study [14] was not as prominent as in the current study. Grauwinkel did however, only use white noise of positive SNRs (i.e., +10 and +5 dB). The current study was part of a larger project in which white noise was...
also included [20]. Results there show no significant differences between white noise and babble noise in their effect on audio-visual voicing perception. In the perception of POA the formant transitions, especially for F₂ and F₃, are important cues [21]. The voice onset and trajectory frequencies of F₂ and F₃ for oral stops are, however, typically above 500 Hz, whereas one of the most prominent differences between babble and white noise is that babble noise has higher energy densities at low frequencies (0-500 Hz). Thus, although voicing differences are susceptible to noise, the obtained voicing differences in the perception of POA are most likely attributed to speech signal characteristics rather than differences between white and babble noise.

Research shows that in an /a/ vowel context, labial, alveolar and velar stop consonants differ in spectral distribution at F₂ and F₁ loci at voice onset [12, 22]. This difference in spectral distribution is, however, more prominent for voiced stops than for voiceless stops [12]. Spectrally, the voiceless /pa/, /ta/ and /ka/ are more narrowly clustered together than the voiced /ba/, /da/ and /ga/ at voice onset. The spectral distributions at voice onset for voiceless stops more closely correspond to the formant frequencies of the following vowel than for voiced stops. The difference in spectral distribution between voiced and voiceless stops may contribute to the differences observed in perception.

Indeed this explanation can at least partly account for the discrepancy found between the current study and that of Behne et al. (2006) using different vowel context. The distinctiveness of voiced and voiceless stops at voice onset is influenced by the vowel context. In an /i/ vowel context, voiced stops become less distinct than voiceless stops, especially at the F₂ locus [12].

The greater spectral distribution of the voice stops compared to the voiceless stops in the /a/ vowel context may render voice stops less susceptible to noise than voiceless stops. Because the F₂ and F₃ of the voiceless stops are distributed across a more restricted set of frequencies than that of the voice stops, they may be more easily obscured by noise at one single relevant frequency. Hence, although voicing is an attribute considered robust in noise it may play an important role in the perception of POA, and thus substantially contribute to the overall saliency of the consonant phoneme in noise. Furthermore, both the current study using babble noise and that of Grauwinkel [14] using white noise showed more auditory responses to voiced than to voiceless POA stimuli. This indicates that the direction of the voice effect on perception of POA is consistent across noise conditions.

5. Conclusions

In AV speech perception POA and voicing cues clearly depend on each other for reliable auditory perception of consonant phonemes. As previous research has shown, visual cues are more likely to be used in speech perception in noise [e.g., 1,2] Furthermore, in an /a/ vowel context, the POA of stop consonants are more auditorily salient when the consonants are voiced compared to when they are voiceless. Although not easily generalizable across vowel contexts, the greater spectral distinctiveness of voiced stops may contribute to their greater auditory prominence relative to voiceless stops. This voicing effect found for babble noise is consistent with that found in a study using white noise [14]. The current study shows, in line with some previous research, that stop consonant POA and voicing are interdependent in /a/ vowel context, and that this interdependency is predictable across different noise type conditions, here shown in babble noise, a finding not previously reported.

6. References
