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

Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

Changes in audio-visual speech perception during adulthood

Dawn Behne[1], Yue Wang[2], Magnus Alm[1], Ingrid Arntsen[1], Ragnhild Eg[1], Ane Valsø[1]

[1] Psychology Department, Norwegian University of Science and Technology, Trondheim Norway
[2] Department of Linguistics, [2]Simon Fraser University, Vancouver BC, Canada

dawn.behne@svt.ntnu.no

Abstract

Audiovisual speech perception research has shown an increasing use of visual information from infancy to young adulthood. The current study extends these findings by examining audiovisual speech perception from young adulthood to mid-adulthood by addressing the extent to which audio, visual and audiovisual cues are used for place of articulation identification. Responses were gathered with young adults (19-30 yrs) and mid-aged adults (49-60 yrs) for voiceless and voiced audiovisual consonant-vowel syllables differing in consonant place of articulation. Materials were presented in quiet and in café noise (SNR=0dB). Results show that mid-aged adults made greater use of visual information than young adults in both quiet and noise, suggesting more than compensation for natural changes in hearing. This was evident across places of articulation and voicing conditions where mid-aged adults showed further indications for using visual cues. Findings indicate that the processing of sensory information continues to change in the course of adulthood with the use of visual cues in audiovisual speech perception increasing with the experience that comes with age.

Index Terms: speech perception, audio, visual, adult development

1. Introduction

1.2. Background

With development from infancy to old age, progressive perceptual learning is accompanied by changes in the peripheral auditory and visual systems, with potential influences on the use of audio, visual and integrated audiovisual cues in audiovisual speech perception.

Speech perception research on development from infancy to young adulthood has shown a general trend of increasing use of visual information (e.g., [1], [2]) and increasing audiovisual integration (e.g., [3], [4], [5]). While this may in part be a result of ongoing peripheral vision development (e.g., [1]), increasing perceptual learning and the associations between auditory and visual information that come with development have also been shown to play a important role (e.g., [6], [4]).

From young adulthood, the auditory and visual systems are fully developed, and with increased age undergo sensory reduction (e.g., [7]). Perceptual learning nevertheless can continue well into old-age and may lead to more efficient use of available cues (e.g., [8]). Brain imaging shows that older adults process information differently than young adults (e.g., [9]) and cross-sectional and longitudinal studies of cognitive aging have shown a change in perceptual and cognitive processes across the adult life span (e.g., [10],[11],[12]). These findings raise the question of how audiovisual speech perception, and in particular, how the use of audio, visual and integrated audiovisual cues changes during adulthood.

In a study comparing AV-fusion by near-normal hearing 18-35 and 65-74 year-olds, Cienkowski and Carney [13] found no difference in AV-fusion responses. Although they did not directly address the extent to which A or V cues may be differentially used by the two groups, they show a general tendency for younger adults to use A cues whereas older adults may tend to make greater use of V cues.

1.3. Current study

The current study extends this research and studies the use of audio, visual and integrated audiovisual information from young adulthood into mid-adulthood. In addition, background noise, consonant voicing and stimulus structure are considered.

1.3.1. Quiet and café noise

Among young adults, background noise (SNR=0dB) leads to more audiovisual fusion responses (e.g., [14, [15]). If changes in peripheral hearing account which for differences in audiovisual perception between younger and older adults, this differences should be reduced in noise, but not in quiet. This is tested in the current study with the use of natural café noise.

1.3.2. Voiceless and voiced consonants.

For young adults, consonant voicing is conveyed efficiently by auditory cues, whereas place of articulation is conveyed via visual cues [16]. In particular, with stimuli incongruent for place of articulation voiced consonants have been widely observed to lead to more audiovisual fused responses than voiceless consonants. The susceptibility of voicing cues to noise interference suggests the potential use for visual cues, in particular for older adults.

1.3.3. Incongruent stimulus structure.

Previous research has shown that for young adults incongruent stimuli with a auditory velar component (A_{velar}) and a visual labial component (V_{labial}) are more likely to lead to audiovisual fused responses than $A_{labial}V_{velar}$ stimuli (e.g., [3]) which, in turn, are more likely to lead to auditory responses. The current study tests the extent to which this pattern holds for older adults.

2. Method

Electrophysiological studies have shown that audiovisual perception is not the same as the combination of single modality sensory processing, such as hearing or vision. (e.g., [17], [18]). In the current study the same incongruent audiovisual stimuli are the basis for testing the use of audio, visual and integrated audio-visual cues.

2.2. Participants

Participants were 10 young adults between 19 and 30 years old (mean=23 yrs) and 10 middle-aged adults between 49 and 60 years old (mean=53 yrs). Each group had a balance between male and female participants, all of which had Norwegian as their native language. All participants reported having normal hearing and normal or corrected-to-normal

2.3. Stimuli

The stimuli were developed from consonant-vowel (CV) audiovisual syllables (/pi/, /bi/, /ti/, /di/, /ki/, /gi/, /pa/, /ba/, /ta/, /da/, /ka/, /ga/) recorded from an adult male native speaker of Norwegian using a Sony mini DV video camera and an external Røde NT3 microphone.

Based on these recordings, the incongruent audiovisual CVs presented in Table 1 were prepared with a labial consonant in one modality and a velar in other modality. The consonants were either voiceless (/p/ or /k/) or voiced (/b/ or /g/), although within any given audiovisual stimulus voicing of the two modalities was the same (e.g., audio /b/ with visual /g/). The vowel was either /i/ or /a/ to allow for differing effects of vowel context (e.g., [19]), but was the same across modalities for a given stimulus.

Table 1: *AV stimuli incongruent for place of articulation*

	$A_{labial}V_{velar}$	$A_{velar}V_{labial}$
Voiceless	pi-ki pa-ka	ki-pi ka-pa
Voiced	bi-gi ba-ga	gi-bi ga-ba

Audiovisual syllables were presented in quiet and unintelligible café noise (0 dB SNR) (e.g., [15]). All CVs were normalized to 70dB. Stimuli in quiet and in café noise were blocked and randomized within each block. With 3 repetitions, each participant was presented 48 stimuli (2 stimulus structures ($A_{labial}V_{velar}$, $A_{velar}V_{labial}$), 2 voicing conditions (voiceless, voiced), 2 background conditions (quiet, noise), and 2 vowels (/i/, /a/)).

2.4. Procedure

Participants were tested in the Psychology Department at the Norwegian University of Science and Technology. AV stimuli were presented on individual 17" computer monitors (1440x900 pixels) at a distance of ca. 50cm in front of each participant and over AKG K271 headphones at ca. 68 dBA.

For each AV syllable a participant´s task was to identify the syllable and give a respond from among syllables with initial /p, t, k, b, d/ and /g/.

3. Results and Discussion

Results presented here focus on the initial consonant in the stimuli. For each stimulus, corresponding responses were tabulated based on whether the consonant in the response matched the consonant in the audio component of the stimulus (A), the video component of the stimulus (V) or was intermediate to the A and V components (AV-fusion.).

An analysis of variance was carried out with stimulus structure ($A_{labial}V_{velar}$, $A_{velar}V_{labial}$), initial consonant voicing (voiceless, voiced), and stimulus background (quiet, café noise) as repeated measures, and age (young adults, mid-aged adults) as a nonrepeated measure.

3.2. Age

The general pattern of A, V and AV-fusion results for age is presented in Figure 1. Whereas young adults generally use A cues more than mid-aged adults [F(1,18)=11.25, p=.004], mid-aged adults use more V cues than young adults [F(1,18)=7.99, p=.011], with no difference in AV-fusion responses between the two groups [F(1,18)=1.51, n.s.].

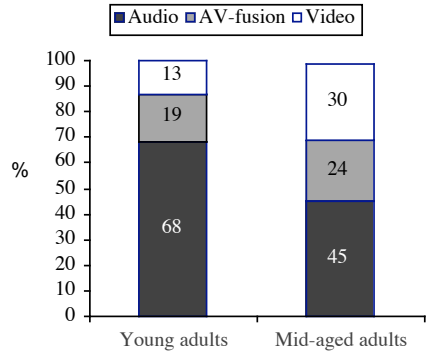

Figure 1: *Percent audio, AV-fusion and video responses by young adults and mid-age adults.*

3.3. Stimuli in quiet and café noise

Results for stimuli in quiet and in café noise are shown in Figure 2. As expected (e.g., [15], [20]), in café noise A cues were generally used less [F(1,18)=38.08, p<.001] and visual cues were used more for both groups of listeners [F(1,18)=17,67, p<.001], although background had no reliable effect on fusion responses [F(1,18)=1.10, *n.s.*].

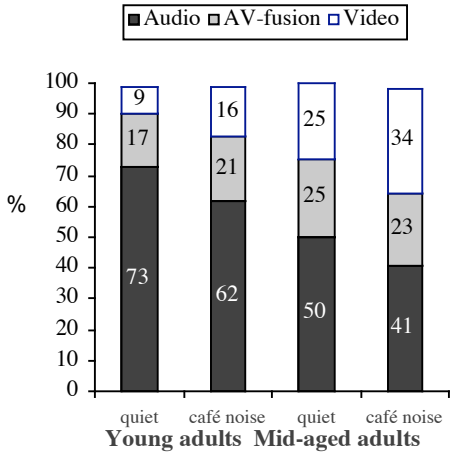

Figure 2: *Percent audio, AV-fusion and video responses in quiet and café noise by young adults and mid-age adults.*

Background café noise was included to neutralize possible differences in auditory acuity for the two groups of listeners. Notably, no interaction was observed between age and background for A [F(1,18)=0.82, n.s.], or V responses [F(1,18)=0.32, n.s.]. That is to say, responses by the two age groups were not differentially affected by the quiet and café noise backgrounds, suggesting no reliable difference in peripheral hearing function with noise between the two groups.

Although not reliable, it should be noted that for fusion responses, a tendency towards an interaction between background and age was observed [F(1,18)=3.43, n.s. (p<.08)], with young adults tending to have fewer fusion responses in quiet than in noise, and a relatively high proportion of fusion responses in both quiet and café noise for

mid-aged adults. This pattern may reflect the additional ca 30 years of experience with integrating AV cues that the mid-aged adults have over the young adults.

3.4. Voiceless and voiced stimuli

Previous research has consistently shown a greater likelihood for AV-fused responses with voiced than voiceless stimuli (e.g., [3]). This is also observed in the current study as is illustrated in Figure 3 [F(1,18)=39.46, p<.001]. In addition, analyses of A and V responses show greater use of A [F(1,18)=14,27, p<.001] and V cues [F(1,18)=8.68, p=.009] for voiceless stimuli than voiced stimuli. That is, although voiced stimuli lead to a greater proportion of fused responses than voiceless stimuli, A and V cues are independently used to identify place of articulation for voiceless stimuli. Furthermore, V cues are especially used in café noise (voiceless, mean=32%; voiced mean=18%) compared with in quiet (voiceless, mean=19%; voiced mean=14%) [F(1,18)=6.82, p=.018]. This pattern is consistent for young and mid-aged adults with no interaction between voice and age for A [F(1,18)=0.57, n.s.], V [F(1,18)=0.04, n.s.], or AV-fusion [F(1,18)=0.54, n.s.] responses.

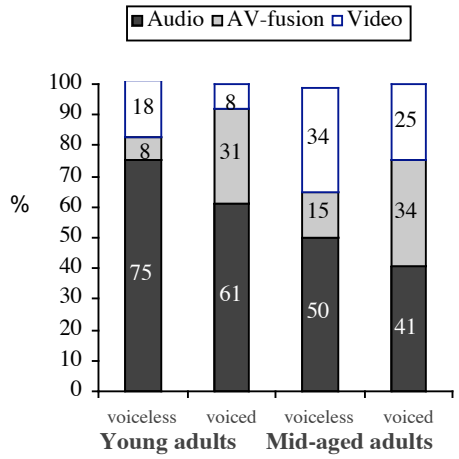

Figure 3: *Percent audio, AV-fusion and video responses for voiceless and voiced initial consonants by young adults and mid-age adults.*

3.5. $A_{labial}V_{velar}$ and $A_{velar}V_{labial}$ stimuli

As is shown in Figure 4, the commonly observed pattern (e.g., [3]) in which an $A_{labial}V_{velar}$ stimuli lead to fewer A responses [F(1,18)=126.03, p<.001] and more AV-fused responses [F(1,18)=112.91, p<.001] than $A_{velar}V_{labial}$ is observed for young and mid-aged adults in the current study. Furthermore, as is also illustrated in Figure 4, an interaction for age and stimlus structure for V cues [F(1,18)=5.17, p=.035] shows that whereas the two age groups had comparable use of V cues for $A_{labial}V_{velar}$, mid-aged adults also made use of V cues for $A_{velar}V_{labial}$ stimuli, a pattern not observed for young adults. As is shown in Figure 5, this pattern occurs in both noise conditions, although to a slightly greater extent in noise than in quiet [F(1,18)=5,58, p=.030].

These findings demonstrate mid-aged adults using the same cues as young adults, but additionally making use of subtle V cues for place of articulation which were not used by young adults.

Furthermore, whereas noise generally is believed to increase the use of visual cues, the results here show that this is only the case if the cue is already in use. In the case of young adults who did not use V cues for $A_{velar}V_{labial}$ stimuli, they continued to use A cues [F(1,18)=13.76, p=.002] and did not turn to using V cues in noise, whereas mid-aged adults who did make use of V cues in quiet made use of them to a greater extent in noise.

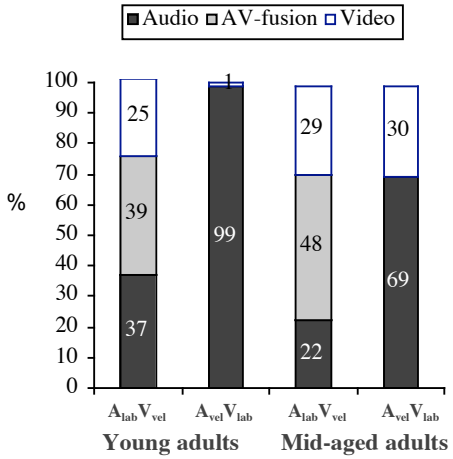

Figure 4: *Percent audio, AV-fusion and video responses for $A_{labial}V_{velar}$ and $A_{velar}V_{labial}$ stimuli by young adults and mid-age adults.*

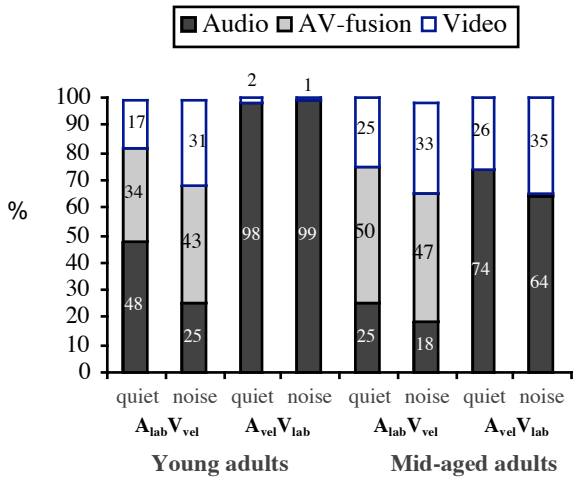

Figure 5: *Percent audio, AV-fusion, and visual responses in quiet and café noise for $A_{labial}V_{velar}$ and $A_{velar}V_{labial}$ stimuli by young adults and mid-age adults.*

3.6. Voiceless and voiced $A_{labial}V_{velar}$ and $A_{velar}V_{labial}$

Further analyses of an interaction between stimulus structure, voicing and age for V cues [F(1,18)=12.17, p=.003], illustrated in Figure 6, shows that for $A_{labial}V_{velar}$ stimuli, mid-aged adults made greater use of V cues when the stimuli were voiceless than when they were voiced. This was in addition to having A [F(1,18)=1.75, n.s] and AV-fusion [F(1,18)=0.72, n.s] responses comparable to young adults. That is to say, the mid-aged adults are again making great use of subtle V cues than the young adults.

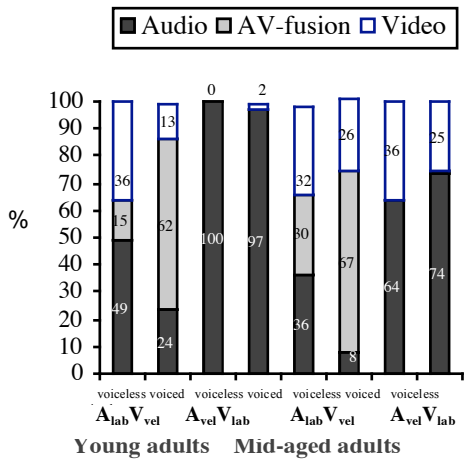

Figure 6: *Percent audio, AV-fusion and video responses in voiceless and voiced $A_{labial}V_{velar}$ and $A_{velar}V_{labial}$ stimuli by young adults and mid-age adults.*

4. Conclusions

The current study addressed the issue of how progressive perceptual learning and changes in the peripheral auditory and visual systems influence the use of audio, visual and integrated audiovisual cues in audiovisual speech perception.

Results revealed two unexpected findings of interest. First, whereas noise generally is believed to increase the use of visual cues, the results here suggest that this is only the case if the cue is already in use. Young adults who did not use certain V cues for place of articulation identification, did not use them in noise, whereas mid-aged adults who did make use of V cues in quiet made use of them to a greater extent in noise. Second, findings support previous research showing a greater likelihood for AV-fused responses with voiced than voiceless stimuli (e.g., [3]), and further show that A and V cues are independently used to identify place of articulation for voiceless stimuli for young and mid-aged adults.

Previous developmental research on infancy to young adulthood suggests a trend toward increased use of visual cues (e.g., [1][2]) and audiovisual integration (e.g., [3], [4], [5]). Background café noise was included to neutralize possible differences in auditory acuity for the two groups of listeners. Notably, responses by the two age groups were not differentially affected by the quiet and café noise backgrounds, suggesting no reliable difference in peripheral hearing function with noise between the two groups. In addition, mid-aged adults used the same cues as young adults for place of articulation identification, and in addition made use of subtle visual cues which were not used by young adults.

Results extend previous research and demonstates a continuated increase in the use of visual cues beyond young adulthood into mid-adulthood. Findings indicate that processing of sensory information continues to change in the course of adulthood, with the use of visual information in audiovisual speech perception robustly increasing with the experience that comes with age.

5. References

[1] Massaro, D., "Children's perception of visual and auditory speech", Child Development, 55: 1777–1788, 1984.

[2] Robinson, C.W., and Sloutsky, V.M., "Auditory dominance and its change in the course of development." Child Development, 75(5): 1387-1401, 2004.

[3] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", Nature, 264: 746-748, 1976.

[4] Rosenblum, L.D., Schmuckler, M.A., and Johnson, J.A., "The McGurk effect in infants." Perception and Psychophysics, 59(3): 347-357, 1997.

[5] Wightman, F., Kistler, D., and Brungart, D., "Informational masking of speech in children: Auditory-visual Integration", J. Acoust. Soc. Am., 119(6): 3944, 2006.

[6] Green, K.P., "Studies of the McGurk Effect: Implications for theories of speech perception." Proceedings of the International Conference on Spoken Language, 3:1652-1655, 1996.

[7] Abel, S. M., and Hay, V. H., "Sound localization: The interaction of HPDs, aging and hearing loss", Scand. Audiol., 25: 3–12, 1996.

[8] Hardison, D.M., "Acquisition of second-language speech: Effects of visual cues, context, and talker variability", Applied Psycholinguistics, 24: 495-522, 2003.

[9] Madden, Dj., Whiting, W.L., Provenzale, J.M., and Huettel, S.A., "Age-related changes in neural activity during visual target detection measured by fMRI", Cereb. Cortex, 14: 143-155, 2004.

[10] Kline, D., and Schieber, F., "Vision and aging". In J. E. Birren and K. W. Schaie (Eds.), Handbook of the Psychology of Aging (pp. 296 –331). New York: Van Nostrand Reinhold, 1985.

[11] Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N., Smith, A. D., and Smith, P., "Models of visuospatial and verbal memory across the adult life span", Psychology and Aging, 17(2): 299-320, 2002.

[12] Salthouse, T. A., "The processing-speed theory of adult age differences in cognition", Psychological review, 103:403-428, 1996.

[13] Cienkowski, K.M. and Carney, A.E., "Auditory-visual speech perception and aging", Ear and Hearing, 23:439-449, 2002.

[14] Sumby, W. and Pollack, I., "Visual contribution to speech intelligibility in noise", J.Acoust.Soc.Am. 26:212-215, 1959.

[15] Sekiyama, K. and Tohkura, Y. "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility", J. Acoust. Soc. Am. 90:1797-1805, 1991.

[16] Grant, K.W., Walden, B.E., and Seitz, P.F., "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration", J. Acoust. Soc. Am., 103:2677-2690, 1998.

[17] Bernstein, L. E., Ponton, C., and Auer, E. T., "Is audiovisual speech integration an early perceptual effect? An event-related potential study of the McGurk effect", Cognitive Neuroscience Society, New York City, March 25-27, 2001.

[18] McPherson, J. L. and Andrews, S. M., "Mismatched Negativity to Auditory and Visual Discrepancies to Three Phonemes" Paper presented at XV IERASG. 2002.

[19] S. Shigeno, "Influence of vowel context on the audio-visual speech perception of voiced stop consonants", Jpn. Psychol. Res., 42:155-167, 2000.

[20] Fixmer, E., and Hawkins, S., "The influence of quality of information on the McGurk Effect", Presented at the

Australian Workshop on Auditory-Visual Speech Processing. 1998.