Visual Speech Information Aids Elderly Adults in Stream Segregation

Alexandra Jesse [1], *Esther Janse* [2,1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Utrecht institute of Linguistics, Utrecht University, Utrecht, The Netherlands
Alexandra.Jesse@mpi.nl, Esther.Janse@mpi.nl

Abstract

Listening to a speaker while hearing another speaker talks is a challenging task for elderly listeners. We show that elderly listeners over the age of 65 with various degrees of age-related hearing loss benefit in this situation from also seeing the speaker they intend to listen to. In a phoneme monitoring task, listeners monitored the speech of a target speaker for either the phoneme /p/ or /k/ while simultaneously hearing a competing speaker. Critically, on some trials, the target speaker was also visible. Elderly listeners benefited in their response times and accuracy levels from seeing the target speaker when monitoring for the less visible /k/, but more so when monitoring for the highly visible /p/. Visual speech therefore aids elderly listeners not only by providing segmental information about the target phoneme, but also by providing more global information that allows for better performance in this adverse listening situation.

Index Terms: speech perception, audiovisual alignment, stream segregation, aging

1 Introduction

Listening to a speaker while others talk can be difficult. The listener has to unravel the mixture of speech streams and focus on the target speaker's speech. This poses a challenge especially when getting older. Elderly adults are more affected by competing speech than young adults [1, 2]. This age difference is larger for hearing a single competing speaker than for other types of background noise [3]. We asked whether elderly listeners can benefit in this situation from seeing the target speaker during speech processing. We further examined what information visual speech provides for this benefit.

When aging, listeners gradually lose some aspects of their hearing acuity. Age-related hearing deficits are due to an overall decline in sensitivity, but also, for example, due to a decrease of temporal and frequency resolution [e.g., 4]. Age-related hearing loss is often the main and sometimes even the sole predictor of various speech perception deficits in the elderly [5, 6, 7]. Age-related hearing loss contributes to the disproportional difficulty elderly listeners face when listening to speech while another speaker talks [2, 8, 9]. This is not surprising, given that a variety of the acoustic cues used to segregate speech streams [e.g., temporal synchrony or pitch; 10, 11] are impacted by age-related hearing loss. Elderly listeners are also impacted in their perception of speech by cognitive aging deficits, especially in adverse listening situations. Elderly adults are more affected by competing speech than young adults even when both groups are matched on hearing acuity [1], suggesting that this difficulty does not increase solely due to age-related hearing loss. Some of the cognitive abilities declining with age are working memory capacity, information processing speed, and inhibitive control [12, 13, 14]. Working memory span and the ability to inhibit irrelevant information, but not information processing speed, predict elderly listener's performance when presented with competing speech [9, 15]. These factors play a role as listeners not only have to separate the speech streams but also have to retain focus on the target speech while inhibiting the processing of the competing speech. Elderly listeners are also, in comparison to young adults, especially affected by the meaningfulness of competing speech [2]. This also suggests an influence of cognitive aging, as an age-related decline of hearing alone cannot predict this result. Furthermore, elderly adults' ability to recall speech is predicted by their executive control abilities when the target speech is presented along with meaningful competing speech [2]. In summary, when elderly adults listen to a speaker while another one talks, their hearing acuity largely governs their ability to comprehend the target speaker. In addition, elderly adults' ability to inhibit the competing speech, their control abilities to maintain focus on the target speech as well as their memory capacities determine differences in their comprehension abilities.

In the present study, we examined whether elderly listeners benefit when they not only hear but also see the target speaker while another speaker is also audible. More specifically, we tested whether elderly adults can use visual speech information sufficiently and rapidly during the processing of a sentence to improve their comprehension. Elderly adults are generally worse at lip-reading than young adults [16, 17], but their lip-reading ability is not related to their hearing sensitivity [18]. Nevertheless, when presented with speech in multispeaker babble noise, elderly and young adults equated on their lip-reading abilities show similar sized audiovisual benefits [16, 17]. Elderly and young adults thus do not seem to differ in their ability to benefit from the extracted visual speech information. These results were obtained, however, in tasks that did not require speeded responses, but rather allowed for unlimited processing time. The audiovisual benefits for elderly and young adults could therefore have resulted from different processing levels. More specifically, the audiovisual benefit observed for elderly adults could have solely emerged during additional post-perceptual processing of the visual speech. Elderly adults may not be able to rapidly cope with the additional demands of processing visual speech during speech perception. Rather, they may need more time to process this additional information and only consider it at later post-perceptual stages of recognition. This seems especially likely in more resource demanding situations, such as when listening to speech while others talk.

It seems therefore possible that elderly adults do not benefit in comprehension, or at least not to the same degree as young adults, from seeing a speaker when asked to give a speeded response. At least when tapping into comprehension with tasks requiring *speeded and continuous verbal responses*, this seems to be the case. When hearing two speakers simultaneously, elderly listeners did not benefit in their

immediate repetition of the target speaker, when seeing the speaker talk [19, 20]. This shadowing task, however, required speech production during listening. That is, listeners had to plan and produce speech while listening and could have suffered from phonological interference from hearing their own speech. Producing speech while listening could have also competed for cognitive resources that were consequently no longer available for speech perception. Recent evidence suggests that one's own silent articulations alter the perception of simultaneously presented auditory speech from another speaker similarly to how seeing this speaker talk would [21]. Thus, the verbal but not the speeded nature of the shadowing responses could have interfered with the processing of visual speech. Some preliminary evidence suggests that elderly listeners can cope with fast processing of visual speech [22]. Elderly listeners' visual speech processing was not disproportionally affected by the presentation rate of the visual speech. Any age-related decline in information processing speed did not affect visual speech processing. The employed task, however, did not assess processing speed by measuring response latencies but rather by manipulating the presentation rate. That is, this study again allowed for unlimited processing time, which could have allowed for the successful processing of faster visual speech.

In the present study, we used a phoneme monitoring task to assess whether elderly listeners benefit in their comprehension from seeing the target speaker when target speaker and a competing speaker are audible. Phoneme monitoring requires a speeded manual response. No verbal response has to be given. Listeners have to indicate by button press as fast and as accurately as possible when they detect the phoneme for which they are asked to monitor the target speech stream. Even though phoneme monitoring requires a speeded response, the participants are instructed to simultaneously maximize both accuracy and speed of their responses. Phoneme monitoring responses can reflect lexical processing [see 23 for an overview]. Responses result from a race of a prelexical processing route providing phonetic information and a lexical route providing lexical phonological knowledge [24]. When monitoring for phonemes in meaningful sentences, participants seem to rely on the lexical route and hence, responses reflect lexical processing [25]. Importantly, monitoring does not interfere with processing of the to-be-monitored sentences for meaning [26, 27]. Phoneme monitoring thus allows to test whether elderly listeners benefit during speech processing from seeing the speaker. If elderly listeners benefit in their comprehension from seeing the target speaker, then target phonemes should be detected more accurately and more rapidly when the speaker can also be seen.

We further investigated what kind of information visual speech provides by varying the visibility of the target phoneme. Participants had to either monitor for the visually distinct phoneme /p/ or for the visually less distinct phoneme /k/. If visual speech only provides local segmental information about the target phoneme, then the audiovisual benefit should only be found for /p/ and not for /k/. If visual speech aids the listener more globally, that is by providing information to segregate the two streams and attend to the target speaker, then the audiovisual benefit should be found when monitoring for /p/ and for /k/. The audiovisual benefit will be larger for /p/ than for /k/, if both global and local visual speech information aid. To ensure that listeners were not able to respond solely based on visual segmental information alone but were required to combine the target speaker's visual and auditory speech for their response, we also systematically included competitor phonemes on half of the trials that were visually highly confusable with the target phonemes.

2 Experiment

2.1 Participants

Forty native speakers of Dutch over the age of 65 (*M*=72 years, *SD*=5.4 years) participated in the experiment (23 females, 17 males). More than two thirds of them had received higher-level education. Participants with varying degrees of age-related hearing loss were included in the study. Individual hearing losses were determined as the participants' pure-tone average hearing loss over the frequencies of 1, 2, and 4 kHz in their best ear. The average hearing loss was 32 dB (*SD*=12 dB). Only two participants had hearing aids, which they were asked not to wear during the experiment. If needed, participants were asked to wear their appropriate glasses. Participants contacted the researchers in response to an article in a local newspaper and received 10 euros for their participation.

2.2 Stimuli

Visibility of the target phonemes was varied by using the visually highly distinct phoneme /p/ and the visually less distinct phoneme /k/ as targets. Two sets of words were created for each of the target phonemes. Each of these four word sets consisted of 16 monosyllabic and 16 bisyllabic Dutch words that all contained the respective target phonemes only word-initially. These words also did not include any other phoneme from the same viseme group [28] as the target phoneme ({p}=(/p,b,m/); {k}=(/k,r,R,x,ŋ,h/)). All words had primary lexical stress on the first syllable. All four sets were equated on their onset complexity and their average spoken word frequency as taken from the CELEX database for Dutch [29].

The target-bearing words were placed in low cloze-probability sentences of varying length (e.g., "De circusartieste had al jaren een *pil* die haar zenuwen onder controle hield." ["The circus artist took for years a *pill* that kept her nerves under control."]). Sentences varied with regard to the position of the target-bearing word within the sentence. For each target phoneme, words from one of the sets were placed in sentences that did not contain any phoneme from the viseme class of the target phoneme. The words from the other set for each target phoneme were placed in sentences that also contained one word with a viseme competitor in word-initial position. For /p/ targets, this viseme competitor phoneme was /m/; for /k/ targets, the competitor was /x/. The competitor-bearing words always preceded the targets distant enough to distinguish responses to these competitors from responses to targets. The inclusion of these visual competitors ensured that listeners had to use both auditory and visual information to detect targets and could not simply monitor the visual speech stream.

Two sets of foil sentences were created for each target phoneme. Foil sentences did not contain the respective target phoneme. One set of sentences for each target phoneme also did not contain any viseme competitors. The other sets contained one competitor-bearing word. The occurrence of viseme competitors was hence not predictive of whether or not the sentence contained a target. In addition, four practice sentences for each target phoneme were created, where two of them contained a target. One of each foil and target practice trial for a given phoneme also contained a viseme competitor.

All of these sentences were video recorded as spoken by a young female native speaker of Dutch. In these recordings, the main sentence-level accent never fell onto the target-bearing words. The target speaker was also recorded giving instructions for the task. This video was later used to familiarize participants with the speaker before the experiment. Another female Dutch speaker close in age to the target speaker was recorded as distractor speaker. Distractor sentences did not contain the respective target phoneme. Distractor sentences were cut to match in duration the respective foil or target sentence they were assigned to. The amplitude of each cut distractor sentence was modified relative to the amplitude of its assigned target speaker sentence to obtain a signal-to-noise ratio of +2dB. The modified distractor sentence was added as an audio track to the target speaker video in Adobe Premiere. The onset and offset of a distractor sentence were temporally aligned with the onset and offset of a target speaker sentence. During the subsequent export as a stereo video, Adobe Premiere mixed the two audio tracks and copied this mixed track onto each stereo channel of the video. That is, the mixed audio track of both speakers was presented diotically to the listener. Final videos were converted to the mpg format with audio tracks sampled down from 48 kHz to 32 kHz. Target phoneme onset times were determined based on the acoustic onset in the audio channel of the final video. On auditory-only presentation trials, the same videos were presented as during audiovisual trials, but here a black rectangle covered the video display completely. All videos had a size of 720 by 576 pixels.

2.3 Procedure and design

The experiment consisted of three parts. First, participants were familiarized with the target speaker by watching and listening to an approximately 40 sec long video of the speaker explaining the task. The distractor speaker was not presented during this familiarization. All audio materials in the experiment were presented over headphones at a fixed listening level. Videos were shown on a computer monitor at approximately 50 cm in front of the participants.

Next, participants received two blocks of test trials, one for each of the target phonemes. Within each block, participants monitored for one type of target phoneme only. Each test block was preceded by a practice block, consisting of four practice trials for the respective target phoneme. Participants were asked to monitor the speech of the target speaker for the target phoneme while ignoring the competing speaker. They were instructed to always watch the screen, as on some trials, the target speaker was also visible. The distractor speaker would never be visible. Both target and distractor speaker were, however, always audible simultaneously in both ears. Participants were to indicate as fast and as accurately as possible by press of a key on the button box when they perceived a word in the target speech stream that began with the target phoneme. If a sentence did not contain the target phoneme, no response was to be given. Each trial began with a presentation of the target phoneme printed on the center of the screen for one second followed by a black screen for 630 ms. Then, a fixation cross in centered position was displayed for 250 ms. After 500 ms, the video started with the next retrace of the screen. On auditory-only trials, the video display was occluded by a black rectangle. Responses were collected up to 1500 ms after each video's offset. Independent of a response, the video was played completely on each trial. No feedback was given. The inter-trial interval was 50ms.

The order of test blocks was counterbalanced across participants. The order of trials within a block was randomized. Each block consisted of 64 trials containing a target phoneme and 64 foils. 32 of the target and 32 of the foil trials contained viseme competitors. Half of each of these four trial types were presented only auditorily; on the other half, the target speaker was also visible. The assignment of a sentence to modality condition was pseudo-randomized but counterbalanced across participants. Participants took a break in between blocks. The experiment lasted approximately one hour.

2.4 Analyses

Mixed effect models were implemented using the lmer function in the lme4 package [30] in the R statistical program. All responses given after acoustic target onset and within 2.5 standard deviations of their mean (M =2596 ms after acoustic target onset) were considered as correct detections (hits). Models were developed separately to predict performance as measured by hits (i.e., correct target detection) and by the log-transformed response latencies of these hits. Given the categorical nature of hits, a binomial logit linking function between hits and predictors was included into these models [31]. P-values for the log response latency models were calculated based on Markov chain Monte Carlo simulations (n =10,000) with R's pvals.fnc function. Systematic step-wise model comparisons using likelihood ratio tests established the best-fitting model. Modality (auditory-only, audiovisual), target phoneme (i.e., target visibility; /p/, /k/), and competitor presence (present, absent) were evaluated as categorical fixed predictor variables. In addition, block (two levels) was evaluated as a categorical control variable. For categorical fixed factors, one condition is mapped onto the intercept of the model. The model estimates the degree to which the intercept has to be adjusted to account for performance observed under another condition of the factor. If the adjustment is significantly different from zero, the factor has a significant effect on performance. For the categorical factors considered here, the auditory-only condition for target phoneme /k/ in block 1 with no preceding visual competitor was mapped onto the intercept. We also evaluated target time, that is, when in a sentence a target occurred, and trial as continuous control factors. To infer an effect of a continuous factor, the model evaluates whether an estimated adjustment of the regression slope for this factor differs significantly from zero. All best-fitting models included both subject and item as random factors. This allows the models to make specific adjustments to the regression weights based on the subject's or item's mean.

2.5 Results and discussion

Figure 1 shows average response latencies for both modality conditions for each target phoneme. Responses on audiovisual trials were faster than on auditory-only trials, and this benefit was larger for the monitoring of /p/ (M_A=966 ms, M_{AV}=868 ms) than of /k/ (M_A=934 ms, M_{AV}=895 ms).

The overall best-fitting model explaining log response latencies contained modality and target phoneme, as well as their interaction, as predictors. Competitor presence was not included in the final model, as it did not contribute to a better fit of the model. Block was included as a predictor and allowed to interact with modality condition. Target time and trial also contributed to a better-fitting model. Modality condition had a significant effect on response latencies (β=-.118, p<.00001). Responses were faster for audiovisual than

for auditory-only presentations. This audiovisual benefit varied, however, as a function of target phoneme and block. The audiovisual benefit was larger for /p/ than for /k/ responses (β=-.101, p<.00001). Overall, there was no difference in performance depending on target phoneme (β=.048, p=.11). Regardless of the target phoneme, the audiovisual benefit decreased over blocks (β=.074, p<.0024). Block, however, had no overall effect on performance (β=-.015, p=.41). The overall increase of response latencies over the course of the experiment was better captured by a gradual change over trials (β=.0007, p<.00001).

When a target occurred in a sentence also affected response latencies. The later in a sentence target phonemes were presented, the faster participants detected them (β=-.0004, p<.0024). One of the reasons for this effect is that, the more of the sentence unfolds, the more its semantic content predicts the occurrence of a target-bearing word [27]. This suggests that listeners in the present study also processed the speech materials for meaning. Note, however, that the effect of target position in the sentence did not vary across phoneme types. That is, the target sentences for /p/ and /k/ did not differ in their predictability of the target phonemes. This was further confirmed by a separate analysis for auditory-only trials where no effect of target phoneme was found (β=.046, p=.12) and an interaction of target time with target phoneme did not contribute to a better-fitting model. Target time had an overall effect on performance in auditory-only trials (β=-.00004, p<.014).

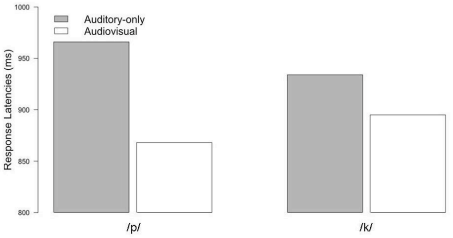

Figure 1: Mean response latencies to /p/ and /k/ in audiovisual and auditory-only trials.

Given that the audiovisual benefit varied in size as a function of target phoneme, we also assessed performance separately for each target phoneme. For the /p/ condition, the final model contained modality and block and their interaction, as well as competitor presence, target time, and trial as fixed factors. Again, responses were faster in audiovisual than in auditory-only presentations (β=-.228, p<.00001). This audiovisual benefit decreased over blocks (β=.100, p<.0037). Overall performance did not vary across blocks (β=-.020, p=.83), but was better captured in a increase of response latencies over trials (β=.0009, p<.003). Responses were faster the later targets were presented in a sentence (β=-.00004, p<.0301). Unlike in the overall model, the performance on /p/-trials was affected by the presence or absence of a visual competitor in the preceding part of the sentence. Responses to target phonemes were slower when a visual competitor had already been encountered in the sentence (β=.074, p<.0407). Seeing a visual competitor slowed down participants' responses to the later occurring target phoneme. For the /k/ condition, the final model only contained modality and trial as fixed factors. There was a significant audiovisual benefit for responses to /k/ (β=-.084, p<.00001). Overall, responses became slower over trials (β=.0005, p<.0329).

Figure 2 shows the average percentage of correct target detections for both modality presentation conditions separately for each target phoneme. Correct detection of both target phonemes improved when the target speaker was also visible. This benefit was larger when monitoring for /p/ (M_A=55%, M_{AV}=81%) than for /k/ (M_A=58%, M_{AV}=70%).

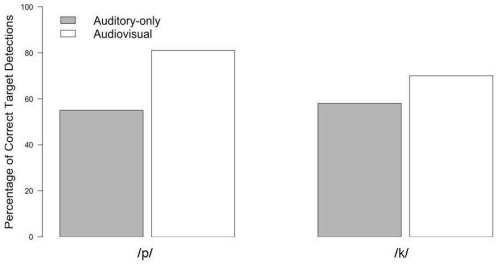

Figure 2: Mean percentage of correct target detection of /p/ and /k/ in audiovisual and auditory-only trials.

The best-fitting model for correct target detection contained modality, target phoneme, target time, and trial as fixed factors. It also allowed for interactions of modality with target phoneme and with target time, respectively. Competitor presence or block did not contribute to a better-fitting model. More target phonemes were accurately detected, when participants not only heard but also saw the target speaker (β=.744, p<.00001). This audiovisual benefit was larger when monitoring for /p/ than for /k/ (β=1.023, p<.00001). The audiovisual benefit decreased the later a target occurred in a sentence (β=-.193, p<.0106). Generally, performance decreased with later trials in the experiment (β=-.003, p<.0162). Overall, there was no difference in performance as a function of target phoneme (β=-.161, p=.45). The later a target was presented in a target, the more likely it was detected (β=.457, p<.00001). Note, however, that the interaction between target phoneme and position of the target-bearing word did not contribute to a better-fitting model, suggesting again, that the target-bearing sentences did not differ in their target predictability across the two phonemes. An additional analysis of auditory-only performance confirmed that auditory detection did not vary as a function of target phoneme (β=-.19, p =.40). There was also no interaction between target time and target phoneme (β=.099, p=.67). Target time had an overall effect on auditory-only performance (β=.44, p<.007).

Since the size of the audiovisual benefit varied across target phonemes, we also assessed performance separately for each target phoneme. For both target phonemes, the best-fitting models contained only modality and target time as fixed factors. The model for /k/ also allowed these two factors to interact. The correct detection of /p/ was more likely for audiovisual than for auditory-only presentations (β=1.849, p<.00001). Target detection was also more likely the later the target phoneme occurred in a sentence (β=.479, p<.0004). When monitoring for /k/, correct detection was influenced by presentation modality. Targets were more likely to be recognized when the target speaker was presented audiovisually than auditory-only (β=.761, p<.00001). /k/-targets were also more likely detected the later they occurred in a sentence (β=.417, p<.0106). The audiovisual benefit was smaller the later a target occurred in a sentence (β=-.253, p<.0133).

3 General Discussion

Elderly listeners are, in comparison to young listeners, disproportionally more affected in their comprehension of a target speaker when hearing a competing speaker [3]. Our study provides evidence that elderly adults benefit in this situation from also seeing the target speaker. This also shows that elderly adults can benefit from visual speech in tasks requiring fast responses. In other words, elderly adults can process the auditory streams and the supplied visual target speech information rapidly and efficiently enough to benefit in their comprehension as speech unfolds.

Previous failures to find an audiovisual benefit for elderly listeners with the shadowing task in stream segregation situations thus seem to be due to task-specific requirements [19, 20]. Most likely, elderly listeners' own productions of speech interfered directly with their perception and/or affected performance by subsuming cognitive capacities needed for audiovisual comprehension. This is in line with the recent finding that one's own produced articulations affect the processing of auditory speech produced by another speaker [21]. Note, however, that an audiovisual benefit for shadowing can be found for young adults [32], even in the more resource demanding situations of processing speech while simultaneously hearing competing speech [e.g., 33, 34]. No benefit is found, however, when more cognitive demands are added, such as when simultaneously translating the to-be-shadowed speech into another language [35].

The ability to benefit from seeing a speaker in stream segregation situations emerges at an early age. When hearing a target and a competing speaker, 7.5-month-old infants were only able to segment continuous target speech into words when seeing the target speaker [36]. This benefit was also found when presented with an oscilloscopic representation of the lip movements of the target speaker. It is not clear, however, whether the visual speech provided information that aided in attending to the target speaker or in segmenting speech.

Our study suggests that visual speech can help stream segregation in several ways. An audiovisual benefit was found in responses latencies and in detection rates when participants had to monitor for the highly visible phoneme /p/ and for the visually less distinct phoneme /k/. The audiovisual benefit was, however, larger when monitoring for /p/ than for /k/, even though there was no such difference for auditory-only presentations. Based on these results, two conclusions can be made about the information provided by visual speech to aid comprehension here. First, given the larger audiovisual benefit for the more visible phoneme /p/, visual speech aids by providing local segmental information about the monitored phoneme. Seeing the speaker enables the participant to be more likely to detect the phoneme, but also to detect the phoneme earlier. This is in line with the phoneme identification results obtained in gating tasks [37, 38]. In gating, increasingly longer parts of the signal are presented to the participant for identification. Identification itself does not require a speeded response. Results from these audiovisual gating studies for Dutch and English suggest that when a speaker can also be seen producing a /p/, the phoneme can be recognized with less of the speech signal provided. Our present study expands these results by showing that this audiovisual recognition benefit holds for a speeded response task where processing time is limited and therefore seems unlikely to be due to post-perceptual processing. Furthermore, we showed that seeing the speaker not only aids correct phoneme recognition but also benefits the recognition speed.

Secondly, an audiovisual benefit for response latencies and detection rates was also found for /k/. Thus the audiovisual benefit in this experiment was not entirely due to local segmental information about the target phoneme itself, as /k/ is visually not very distinct [28]. Rather, the audiovisual benefit for detecting /k/ seems to be due to visual information in the carrier sentence preceding the target phoneme. The amount of preceding visual carrier information seems not to be critical. The audiovisual benefit for response speed did not change with the amount of preceding visual speech for /k/ responses. For correctly detecting /k/ phonemes, however, the audiovisual benefit decreased with more preceding context. This decrease seems to be an artifact of detection rates approaching ceiling level for later-occurring targets in auditory-only presentations and thus leaving less room for improvement.

Future research will have to clarify further how visual speech aids comprehension in this task. Visual speech could aid performance at various processing levels. Visual speech could help directly with the segregation of the speech streams, by providing, for example, dynamical information highlighting the temporal synchrony between target auditory speech and visual motion. Similarly, visual speech could (also) help retaining the attentional focus on the target speaker's speech. Last, visual speech could aid the comprehension of the segregated and attended speech stream by providing segmental and prosodic information about the sentence preceding the target. It remains yet to be seen at which of these processing levels visual speech helps comprehension.

4 Conclusions

The present study provides evidence that elderly adults benefit from seeing a speaker when simultaneously hearing the speaker and a competing speaker. Visual speech aids therefore with an important challenge elderly listener encounter. Importantly, this also demonstrates that elderly listeners indeed benefit from visual speech information in their comprehension of speech as it unfolds. Future research has yet to determine how and at which processing stage visual speech benefits the elderly listeners. To understand the contribution of visual speech in this situation more fully, it will be essential to determine what cognitive and perceptual abilities determine the benefit elderly listeners obtain from seeing the speaker.

5 Acknowledgements

This work was supported in part by Innovational Research Incentive Scheme Veni grants from the Netherlands Organization for Scientific Research (NWO) awarded to the two authors, respectively. The authors thank Inge van de Sande and Patrick van der Zande for their help with conducting the experiment and Marieke Pompe and Vera Hoskam for help with the preparation of the materials.

6 References

[1] J. R. Dubno, D. D. Dirks, and D. E. Morgan, "Effects of age and mid hearing loss on speech recognition in noise", *J. Acoust. Soc. Am.*, vol. 76, pp. 87-96, 1984.

[2] P. A. Tun, G. O'Kane, and A. Wingfield, "Distraction by competing speech in young and older adult listeners", *Psychol. and Aging*, vol. 17, pp. 453-467, 2002.

[3] P. A. Tun and A. Wingfield, "Speech recall under heavy load conditions: Age, predictability, and limits on dual-

task interference", *Age and Cognition*, vol. 1, pp. 29-44, 1999.

[4] L. E. Humes and L. Christopherson, "Speech-identification difficulties of hearing-impaired elderly persons: the contributions of auditory- processing deficits", *J. Speech Hear. Res.*, vol. 34, pp. 686-693, 1991.

[5] L. E. Humes, B. U. Watson, L. A. Christensen, C. G. Cokely, D. C. Halling, and L. Lee, "Factors associated with individual differences in clinical measures of speech recognition among the elderly", *J. Speech Hear. Res.*, vol. 37, pp. 465-474, 1994.

[6] J. C. van Rooij and R. Plomp, "Auditive and cognitive factors in speech perception by elderly listeners. III: Additional data and final discussion", *J. Acoust. Soc. Am.*, vol. 91, pp. 1028-1033, 1992.

[7] V. Summers and M. R. Leek, "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss", *J. Speech Lang. Hear. Res.*, vol. 41, pp. 1294-1306, 1998.

[8] D. R. Murphy, J. M. McDowd, and K. A. Wilcox, "Inhibition and aging: Similarities between younger and older adults revealed by the processing of unattended information", *Psychol. and Aging*, vol. 14, pp. 44-59, 1999.

[9] E. Janse, "Hearing and cognitive measures predict elderly listeners' difficulty ignoring competing speech", in *Proceedings of the NAG/DAGA International Conference on Acoustics*, to appear.

[10] G. L. Dannenbring, and A. S. Bregman, "Streaming vs. fusion of sinusoidal components of complex tones", *Perc. and Psychophys.*, vol. 24, pp. 369-376,1978.

[11] J. P. L. Brokx, and S. G. Nooteboom, "Intonation and the perceptual separation of simultaneous voices", *J. of Phonetics*, vol. 10, pp. 23-36, 1982.

[12] F. I. M. Craik and T. Salthouse, *The Handbook of Aging and Cognition*, Lawrence Erlbaum, 2000.

[13] L. Hasher, and R. T. Zacks, "Working memory, comprehension, and aging: A review and new view", *The Psycho. of Learning and Motivation*, vol. 22, pp. 193-225, 1988.

[14] T. A. Salthouse, and E. J. Meinz, "Aging, inhibition, working memory, and speed", *J. of Gerontology: Psych. Sci*., vol. 50, pp. 297-306, 1995.

[15] L. E. Humes, J. H. Lee, and M. P. Coughlin, "Auditory measures of selective and divided attention in young and older adults using single-talker competition", *J. Acoust. Soc. Am.*, vol. 120, pp. 2926-2937, 2006.

[16] K. M. Cienkowski and A. E. Carney, "Auditory-visual speech perception and aging", *Ear and Hear.*, vol. 23, pp. 439-449, 2002.

[17] M. S. Sommers, N. Tye-Murray, and B. Spehar, "Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults", *Ear and Hear.*, vol. 26, 263-275 , 2005.

[18] N. Tye-Murray, M. S. Sommers, and B. Spehar, "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing", *Ear and Hear.*, vol. 28, 656-668, 2007.

[19] L. A. Thompson, E. Garcia, and D. Malloy, "Reliance on visible speech cues during multimodal language processing: Individual and age differences", *Exp. Aging Res.* , vol. 33, pp. 373-397, 2007.

[20] L. A. Thompson and F. A. Guzman, "Some limits on encoding visible speech and gestures using a dichotic shadowing task", *J. of Gerontology: Psych. Sci*., vol. 54B, pp. 347-349, 1999.

[21] M. Sams, R. Möttönen, and T. Sihvonen, "Seeing and hearing others and oneself talk", *Cogn. Brain Res.*, vol. 23, 429-435, 2005.

[22] B. Spehar, N. Tye-Murray, and M. S. Sommers, "Time-compressed visual speech and age: A first report", *Ear and Hear*., vol. 25, pp. 565-572, 2007.

[23] C. M. Connine and D. Titone, "Phoneme monitoring", *Lang. Cogn. Proc.*, vol. 11, pp. 635-645, 1996

[24] A. Cutler and D. Norris, "Monitoring sentence comprehension", in *Sentence Processing: Psycholinguistic Studies presented to Merrill Garrett*, W. E. Cooper and E. T. C. Walker, Eds., Hillsdale, NJ: Erlbaum, 1979, pp.113-134.

[25] A. Cutler, J. Mehler, D. Norris, and J. Segui, "Phoneme identification and the lexicon", *Cog. Psych.*, vol. 19, pp. 141-177, 1987.

[26] H. Brunner and D. B. Pisoni, "Some effect of perceptual load on spoken comprehension", *J. Verbal Learn. Verbal Behav.*, vol. 21, pp. 186-195, 1982.

[27] D. J. Foss and M. A. Blank, "Identifying the speech code", *Cog. Psych.*, vol. 12, pp. 1-31, 1980.

[28] N. van Son, T. M. I. Huiskamp, A. J. Bosman, and G. F. Smoorenburg, "Viseme classifications of Dutch consonants and vowels", *J. Acoust. Soc. Am.*, vol. 96, pp. 1341-1355,1994.

[29] H. Baayen, R. Piepenbrock, and L. Gulikers, *The CELEX Lexical Database* [CD-ROM], Philadelphia, PA: Linguistic Data Consortium, Univ. of Pennsylvania, 1995.

[30] D. M. Bates and D. Sarkar, *lme4: Linear mixed-effects models using s4 classes*, R package version 0.999375-27.

[31] T. F. Jaeger, "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models", J. Memory and Lang., vol. 59, pp. 434-336, 2008.

[32] D. Reisberg, J. McLean, and A. Goldfield, "Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli", in *Hearing by eye: The psychology of lip reading*, B. Dodd and R. Campbell, Eds., London: Lawrence Erlbaum Associates, 1987, pp. 97-113.

[33] J. Driver and C. J. Spence, "Covert spatial orienting in audition, Exogenous and endogenous mechanisms", J. Exp. Psych.: Human Perc. and Perf., vol. 20, pp. 555-574, 1994.

[34] M. Vitkovitch and P. Barber, "Effect of video frame rate on subjects' ability to shadow one of two competing verbal passages", *J. Speech Lang. Hear. Res.*, vol. 37, pp. 1204-1210, 1994.

[35] A. Jesse, N. Vrignaud, M. M. Cohen, and D. W. Massaro, "The processing of information from multiple sources in simultaneous interpreting", *Interpreting*, vol. 5, pp. 95-115, 2000.

[36] G. Hollich, R. Newman, and P. Jusczyk, "Infants use of synchronized visual information to separate streams of speech", Child Develop., vol. 76, pp. 598-613, 2005.

[37] P. M. T. Smeele, "*Perceiving speech: Integrating auditory and visual speech*", Ph.D. dissertation, University of Technology, Delft, The Netherlands, 1994.

[38] A. Jesse, "*Towards a lexical fuzzy logical model of perception: The time-course of information in lexical identification of face-to-face speech*", Ph.D. dissertation, University of California, Santa Cruz, 2005.