Supplemental Material S1. Descriptive statistics for the sensory measures obtained in Experiments 1 and 2, and cognitive measures obtained in Experiment 2. Exploratory analyses examining how these measures relate to the presentation conditions are also presented. Keyword recognition data for Experiments 1 and 2 are available from https://osf.io/h9p2w/?view_only=81fff81390f741c19a098586b964e448.

Experiment 1

Model Comparison

The initial generalized linear mixed model used in Experiment 1 examined the interaction between Presentation condition (Static; One talking, Two talking, Four talking, and Six talking faces) and Age group (YA, OA). A comparison of this model (interaction with Age group and Presentation condition) with one without Age group (i.e., only Presentation condition), showed a statistically significant difference, $\chi^2(5) = 23.87$, $p = .0002$. The performance scores (from the r performance package, Lüdecke et al., 2021) indicated the model of the interaction between Age group and presentation condition had a higher performance score (57.14% vs. 42.86%). Note, the performance score represents a composite index of AIC, BIC, R2, ICC, RMSE and Log_loss (sigma was not used in this case) and ranges from 0% to 100%, higher values indicating better model performance. Also, score value do not necessarily sum up to 100% as calculation is based on normalizing all indices (i.e., rescaling them to a range from 0 to 1), and taking the mean value for each model.

A similar result was obtained when comparing the model of the first keyword data that included the interaction between Age group and Presentation condition vs. one that included Presentation condition only, i.e., there was a significant difference between the models, $\chi^2(9) = 40.86$, $p = .000005$, with the interaction model having a higher performance score, 62.50% vs. 37.50%. Likewise with the models of last keyword data, there was a difference between the interaction model (Age group and Presentation condition) vs. the Presentation condition only model, $\chi^2(9) = 33.24$, $p = .00001$, with the interaction once again having a higher performance score, 62.50% vs. 37.50%.

Visual Acuity

Younger Adults

All younger participants had normal or corrected to normal vision (i.e., ≥ 1.0 on the FrACT visual acuity measure; Bach, 2007). Younger adults' visual acuity scores ranged from .99 to the maximum score of 2.0 ($M = 1.41$, $SD = .30$). Since younger adults' visual acuity was within the normal range, no analyses were run to determine how visual acuity may have affected recognition performance.

Older Adults

Eight OA had worse than normal vision (i.e., < 1.0 on the FrACT visual acuity measure) with visual acuity scores ranging from 0.71 to the maximum score of 2.0 ($M = 1.21$, $SD = .37$). To determine whether visual acuity affected word recognition performance for the various presentation conditions for the OA, a GLMM (estimated using ML and BOBYQA optimizer) was fitted to predict OA's recognition score for the interaction of Presentation condition and Visual acuity (formula: score ~ Presentation condition × Visual acuity). The model included participant and item (sent) as random effects (formula: list(~1 | participant, ~1 | sent)). The model's total explanatory power (conditional R^2) was 0.42, (the fixed effects marginal R^2 =

0.15). There was a significant effect of Presentation condition ($p < .0.001$); the effect of Visual acuity was not significant ($p = .6$), and neither was the interaction between Presentation condition and Visual acuity ($p = .43$). Given that the main effect of Visual acuity was not significant, no further analyses were conducted.

Hearing Sensitivity

A summary of the hearing level data for younger and older adults is shown in Table 4 (Experiment 1). All younger participants had normal hearing (i.e., ≤ 25 dB HL at 0.25, 0.5, 1, 2, 4 kHz). Older adults' hearing levels were more varied, ranging from normal to moderate-severe hearing loss (i.e., ≥ 40 dB and ≤ 70 dB HL at one frequency), with the majority of older participants (i.e., 14) having only mild hearing loss (i.e., ≥ 25 dB and ≤ 40 dB HL for all tested frequencies).

Table S1. Hearing levels for younger and older adults.

		Experiment 1		Experiment 2	
Hearing Level	**Definition**	**Younger (*n* = 24)**	**Older (*n* = 24)**	**Younger (*n* = 20)**	**Older (*n* = 20)**
Normal	≤ 25[1] at all frequencies[2]	24	7	20	4
Mild Loss	>25 – ≤ 40 at one frequency	0	13	0	9
Moderate Loss	> 40 – ≤ 55 at one frequency	0	2	0	4
Moderate-Severe Loss	> 55 – ≤ 70 at one frequency	0	2	0	3

Note. Hearing level definitions adapted from Wayne et al., 2016, and are measured from the better ear.
[1]dB Hearing Loss. [2]All frequencies refers to 0.25, 0.5, 1, 2, 4 kHz.

Mean pure-tone hearing thresholds for each tested frequency are shown in Table 5 (Experiment 1). As can be seen, younger adults had lower thresholds than older adults for both ears at all tested frequencies.

Table S2. Mean pure-tone hearing thresholds for Experiments 1 and 2.

		Experiment 1		Experiment 2	
Ear	**Frequency (kHz)**	***M* dB HL (*SD*)**		***M* dB HL (*SD*)**	
		Younger (*n* = 24)	**Older (*n* = 24)**	**Younger (*n* = 20)**	**Older (*n* = 20)**
Right	0.25	18.33 (4.34)	22.08 (4.40)	16.75 (2.94)	21.25 (5.82)
	0.50	17.50 (5.52)	23.54 (6.51)	14.00 (3.48)	23.50 (7.45)
	1.00	16.04 (4.66)	22.29 (5.89)	13.50 (3.66)	23.75 (9.58)
	2.00	12.71 (4.89)	23.33 (8.68)	13.25 (4.95)	28.00 (10.31)
	4.00	9.58 (4.87)	32.50 (14.45)	9.00 (6.20)	37.25 (15.09)
Left	0.25	18.54 (4.54)	20.42 (6.41)	16.50 (2.86)	21.75 (6.13)
	0.50	17.71 (4.42)	21.67 (7.32)	14.25 (4.38)	23.00 (9.65)
	1.00	14.38 (4.96)	22.08 (5.50)	12.25 (3.02)	21.50 (10.27)
	2.00	13.33 (6.54)	22.92 (7.65)	10.75 (5.91)	27.25 (13.13)
	4.00	11.46 (6.51)	34.17 (16.98)	9.00 (7.71)	44.25 (17.72)

Better Ear Average scores were calculated by averaging hearing thresholds across all tested frequencies for each ear and selecting the lower average threshold. The within group variation for the Better Ear Average was greater for older adults (Min. = 13.00, Max. = 34.00, *M* = 23.00, *SD* = 5.79) than younger adults (Min. = 8.00, Max. = 20.00, *M* = 14.00, *SD* = 3.38).

To determine whether the OA's Better ear average hearing level affected word recognition performance for the various presentation conditions, a GLMM (estimated using ML and BOBYQA optimizer) was fitted to predict OA's recognition score for the interaction of Presentation condition and Better ear average hearing level (formula: score ~ Presentation condition × Better ear average). The model included participant and item (sent) as random effects (formula: list(~1 | participant, ~1 | sent)). The model's conditional R^2 was 0.43 and the fixed effects marginal R^2 = 0.18. There was a significant effect of Presentation condition (p = .003) and also a significant effect for Better ear average (p = .004), the interaction between Presentation condition and Better ear average was not significant (p = .37).

Model Comparison

Hearing Sensitivity. As detailed above, for OA, the hearing sensitivity data was modelled by an interaction between Presentation condition and Better ear average hearing level. Here, we compare this model with a model that only included Presentation condition, the analysis indicated that there was no statistically significant difference between the models, $\chi^2(5) = 9.06$, p = .11.

Experiment 2

Model Comparison

The initial generalized linear mixed model used in Experiment 2 examined the interaction between Age group (YA, OA) and Presentation condition (Static, One talking, Two talking, Four talking and Six talking faces). A comparison of this model (interaction with Age group and Presentation condition) with one without Age group (i.e., only Presentation condition), showed a statistically significant difference, $\chi^2(5) = 43.22$, p = .000. The performance scores of the model of the interaction between Age group and Presentation condition were higher, 71.43% vs. 28.57%.

Visual Acuity

Younger Adults

All younger participants had normal or corrected to normal vision (i.e., ≥ 1.0 on the FRACT visual acuity measure; Bach, 2007). Younger adults' visual acuity scores ranged from 1.22 to the maximum score of 2.0 (*M* = 1.64, *SD* = .25). Since the range of visual acuity scores was limited, an analysis of whether visual acuity interacted with presentation condition was not carried out for the YA.

Older Adults

Six older adults had worse than normal vision (i.e., a score < 1.0 on the FrACT visual acuity measure), with visual acuity scores ranging from 0.83 to the maximum score of 2.0 (*M* = 1.12, *SD* = .22). As in Experiment 1, to determine whether OA's visual acuity predicted recognition scores in the Presentation conditions, a GLMM (estimated using ML and BOBYQA optimizer) was fitted to predict OA's recognition score for the interaction of Presentation

condition and Visual acuity (formula: score ~ Presentation condition × Visual acuity). The model included participant and item (sent) as random effects (formula: list(~1 | participant, ~1 | sent)). The fitted model's conditional $R^2 = 0.49$ (the fixed effects, marginal $R^2 = 0.1$). There was a significant effect of Presentation condition ($p = .013$); the effect of Visual acuity on recognition score was not significant ($p = .55$) and the interaction between Presentation condition and Visual acuity was not significant ($p = .7$). Given the lack of a main effect of visual acuity, further analysis was not conducted.

Hearing Sensitivity

Table 2 (Experiment 2) summarises hearing sensitivity levels for both younger and older adults. All younger participants had normal hearing (i.e., ≤ 25dB HL at 0.25, 0.5, 1, 2, 4 kHz). As would be expected, younger adults had lower thresholds than older adults at all frequencies for both ears. Since YA all had normal hearing, an analysis of the interaction of hearing level with presentation condition was not carried out.

Older Adults

As in Experiment 1, older adults' hearing levels ranged from normal to moderately severe hearing loss, with the majority of older adults (i.e., 9) having only mild hearing loss. Mean pure-tone hearing thresholds for each tested frequency are shown in Table 3 (Experiment 2). To determine whether OA's Better Ear Average hearing level predicted recognition scores in the visual speech Presentation conditions compared to the Static face baseline, a GLMM was fitted to examine the interaction of Presentation condition and Better Ear Average (formula: score ~ Presentation condition × Visual acuity; participant and item (sent) were included as random effects (formula: list(~1 | participant, ~1 | sent). The fitted model's conditional $R^2 = 0.49$ (the fixed effects, marginal $R^2 = 0.21$). There was a significant effect of Presentation condition (p-value = .041); the effect of Better ear average hearing level on recognition score was also significant ($p < .001$); the interaction between Presentation condition and Better ear average was not significant ($p = .079$).

To determine whether OA's hearing level played a role in speech recognition such that recognition scores for the Two Talking Faces Condition were closer to those of the One Talking Face Condition with better hearing, an additional GLMM was run that examined only the One and Two Talking Faces conditions as a function of Better ear average hearing levels, The results indicated that there was a significant effect of Presentation condition ($p = .002$); the effect of Better ear average hearing level on recognition score was also significant ($p < .001$); the interaction between Presentation condition and Better ear average was not significant ($p = .12$).

Cognitive Tasks

The above approach to analysing the role (if any) of perceptual test scores on speech recognition scores as a function of display condition, was used for the cognitive measures (i.e., to determine whether there are main effects and then examine scores in the One and Two faces conditions as a function of the cognitive test scores).

LSPAN

Younger adults (Min. = 7.00, Max. = 57.00, $M = 25.55$, $SE = 3.47$) scored higher on the listening span (i.e., LSPAN) than older adults (Min. = 0.00, Max. = 23.00, $M = 9.30$, $SE = 1.77$). A GLMM was fitted to examine the interaction of Presentation condition and LSPAN score (formula: score ~ Presentation condition × LSPAN; participant and item (sent) were included as

random effects (formula: list(~1 | participant, ~1 | sent); the fitted model's conditional $R^2 = 0.29$ (the fixed effects, marginal $R^2 = 0.12$). The analysis showed that there was a significant effect of Presentation condition ($p = .01$), and a significant effect of LSPAN ($p < .001$); the interaction between the variables was not significant ($p = .85$).

To determine any relationship between older adults' LSPAN scores and the Presentation conditions, a GLMM was fitted to examine the interaction of Presentation condition and LSPAN score (formula: score ~ Presentation condition × LSPAN; participant and item (sent) were included as random effects (formula: list(~1 | participant, ~1 | sent). The fitted model's conditional $R^2 = 0.49$ (the fixed effects, marginal $R^2 = 0.12$). There was a significant effect of Presentation condition ($p < .001$); the effect of LSPAN on recognition score was not significant ($p = .082$); the interaction between Presentation condition and LSPAN was not significant ($p = .071$).

Trail Making Test (TMT) A and B

Younger adults. To examine whether performance on the TMT-A was associated with the Presentation conditions, a GLMM was conducted to examine the interaction of Presentation condition and TMT-A score (formula: score ~ Presentation condition × TMT-A; participant and item (sent) were included as random effects (formula: list(~1 | participant, ~1 | sent). The fitted model's conditional $R^2 = 0.29$ (the fixed effects, marginal $R^2 = 0.07$). There effect of Presentation condition was not significant ($p = .41$); the effect of TMT-A was significant ($p = .001$) and the interaction between Presentation condition and TMT-A score was not significant ($p = .24$). A similar GLMM was run on the TMT-B scores. The fitted model had a conditional $R^2 = 0.3$ (the fixed effects, marginal $R^2 = 0.07$); the results of the analysis indicated that there was a significant effect of Presentation condition ($p < .001$), but the effect of TMT-B was not significant ($p = .23$).

Older adults. As with the YA, two GLMMs were run, one for the TMT-A scores and the other for the TMT-B ones. The GLMM for the TMT-A scores (formula: score ~ Presentation condition × TMT-A; participant and item (sent) were included as random effects (formula: list(~1 | participant, ~1 | sent) had a conditional $R^2 = 0.49$ (the fixed effects, marginal $R^2 = 0.15$). There was a significant effect of Presentation condition ($p = .021$) and a significant effect of TMT-A score ($p = .026$); the interaction of these effects was not significant ($p = .95$). For the analysis of the TMT-B scores, the GLMM (formula: score ~ Presentation condition × TMT-B; participant and item (sent) were included as random effects (formula: list(~1 | participant, ~1 | sent) had a conditional $R^2 = 0.49$ (the fixed effects, marginal $R^2 = 0.12$). There was a significant effect of Presentation condition ($p = .027$); there was not a significant effect of TMT-B score ($p = .011$); the interaction of these effects was not significant ($p = .99$).