**Supplementary materials**

**Eye-tracking results**

*Native listeners (within-group) - face & mouth*

*Face*

In both clear and degraded speech, native listeners fixated more on the face when no gesture was present (clear speech: CG<CO, *p* = .004, 538 – 1301, degraded speech: DG<DO, *p* = 0.001, 249 – 1461). In line with our predictions, native listeners significantly looked more at the face when speech was degraded than when speech was clear (DO>CO, p <. 001, 1019-2000 ms) (see Figure 6A). When a gesture was present, native speakers also looked more at the face when speech was degraded than when speech was clear (DG>CG, p < 0.001, 1216 – 1936 ms).

*Mouth*

Native listeners significantly gazed more at the mouth when speech was degraded than when speech was clear (DO>CO, p <.001, 1199 – 1990 ms), and when a gesture was present as compared to not present (DG>CG, p < .001, 320 – 1826 ms) (see Figure 7A). It is unclear why these conditions already differed before speech onset, and we speculate this is due to the fact that these time-points are adjacent to the time-interval where the difference between the two conditions is the most prominent, and are therefore clustered together. Moreover, when observing the main effects of Gesture and Noise-Vocoding, this difference is not present. As there is no difference observed for DG vs. CG for non-native speakers over the whole time window (see below), we postulate that this effect reflects that native listeners are less likely to look at the mouth when speech is clear to aid comprehension than non-native listeners. Non-native listeners might aim to extract phonological information for comprehension, possibly because their knowledge of the phonology of the L2 is not as strong as in native listeners (Mayo et al., 1997; Cutler et al., 2004; Bradlow & Alexander, 2007).

 In both clear and degraded speech, listeners gazed more to the mouth when no gesture was present (clear speech: CG<CO, two clusters: *p* = < 0.001, 262 – 377 ms; *p* = .023, 822 – 1344 ms, degraded speech DG < DO, two clusters: *p* = .02, 1059 – 1420, *p* = .019, 1441 – 1551 ms).

*Non-native listeners (within-group - face & mouth)*

*Face*

In line with our hypotheses, non-native listeners looked more at the face when speech was degraded as compared to clear (DO>CO, p < .001, 1243 – 2000 ms), and when a gesture was presented in degraded than in clear speech (DG<CG, p < .008, 1500 – 1883 ms) (see Figure 6B). In both clear and degraded speech, non-native listeners looked more at the face when no gesture was present (clear speech: CG < CO, *p* < .001, 221 – 1378 ms, degraded speech: DG < DO, *p* < .001, 297 – 2000 ms).

*Mouth*

Non-native listeners looked at the mouth more when speech was degraded than when speech was clear (DO>CO, p < .001, 1351 – 2000 ms). However, there was no difference observed between the degraded and clear conditions that contained a gesture (DG = CG, p = .39). In both clear and degraded speech, non-native listeners gazed more often to the mouth when a gesture was not present as compared to present (clear speech: CG<CO, three clusters: *p =* .002, 815 – 1502 ms; *p* = .0026, 1605 – 1830 ms; *p* = 0.03, 1845 – 2000 ms, degraded speech: two clusters: *p* < .001, 191 – 488 ms; *p* = .01, 537 – 1268 ms).

*Native vs. non-native listeners (between-group): face & mouth*

No differences between native and non-native listeners were observed when comparing the differences between the proportions of fixations to the face in the CG-CO (*p* = .35) or DG-CO (*p* = .27) conditions. No differences between native and non-native listeners were observed when comparing the differences between the proportions of fixations to the face in the DG-CG (p = .13) condition. However, the difference between fixations to the face in degraded speech vs. clear speech conditions was larger for native than non-native listeners (Native(DOCO) > Non-native(DOCO), p = .03, 264 – 425 ms).

 We did not observe differences between native and non-native listeners when comparing the differences in fixations to the mouth in the DG-DO (*p* = .16) condition. We did not observe differences between native and non-native listeners when comparing the differences in fixations to the mouth in the DG-CG condition (p = .16), or DO-CO condition (p = .13). However, we observed a larger difference in CG-CO for non-native than native listeners (p = .045, 858 – 992 ms.) indicating that non-native listeners might focus more on other visual articulators than the mouth when speech is clear and a gesture is present than native listeners. We followed up on this by running a post-hoc test and comparing looks to gesture (described below) between the two groups in the CG condition, and confirmed that non-native listeners look more at gestures in clear speech than native listeners (p = .03, 856 - 933 ms).

*Native listeners (within-group) - gesture*

Native listeners looked more at the torso when speech was clear than when speech was degraded (DO<CO, p < .001, 769 – 2001) and when a gesture was present in clear speech than when a gesture was present in degraded speech (DG<CG, p = .002, 1353 – 1884 ms) (see Figure 6C). In both clear and degraded speech, native listeners looked more at the body when a gesture was present than not present (clear speech: CG > CO, *p* < .001, 233 – 1425 ms, degraded speech: DG > DO, *p* < .001, 245 – 2000 ms).

*Non-native listeners (within-group) - gesture*

Non-native listeners initially look more at the body when speech is clear than when speech is degraded (DO<CO, p < .001, 259 – 513 ms), but then look more at the body when speech is degraded as compared to clear (DO>CO, p < .001, 1278 – 2000 ms) (see Figure 6B). When a gesture is present, non-native listeners look more at the body than when speech is clear than when speech is degraded (DG<CG, p = .013, 1521 – 1759 ms). The difference between fixation proportions to conditions that contained a gesture was smaller than the difference between fixation proportions to the body in conditions that did not have a gesture (DGCG < DOCO, p < .001, 1641 – 2000 ms), indicating that similar to native listeners, when non-native listeners process degraded speech in a gesture and no gesture context, visual attention is probably allocated to informational resources for a longer time than when speech is clear.In both clear and degraded speech, non-native listeners look more at the body when a gesture is present (clear speech: CG>CO, *p* < .001, 217 – 1448 ms., degraded speech DG>DO,: *p* = .001, 264 – 2000).

*Native vs. non-native listeners (between-group): gesture*

No differences were observed when comparing gaze to the body when comparing the differences between DO-CO (p = .41) and DG-CG (p = .56) between the native and non-native group. We observed a difference between the groups when comparing the difference DG-DO, which was larger for non-native listeners than native listeners (DGDODE > DGDONL, p = 0.029, 743 - 1087 ms.) and for CG-CO, which was larger for non-native listeners than native listeners (CGCODE > CGCONL, p = .008, 731 - 952), which might indicate that non-native listeners focus more on gestures during both clear and degraded speech as compared to native listeners. In addition to the post-hoc comparison of fixations to CG , we ran another post-hoc test to confirm a similar pattern in degraded speech: non-native listeners indeed significantly look more to gestures in degraded speech than native listeners (p = .01, 981 - 1151 ms).



Figure S1: Differences for significant interactions including the two listener groups. Mean proportion of fixation differences over time on the mouth for native listeners (red) vs non-native listeners (blue) (A/B) and Gesture/Torso (c). In all graphs, the shaded color bars around the mean proportion lines represent standard error. Below each graph, the difference between the conditions and the direction of the effects is specified per comparison. The colored dashed lines represent the differences between the conditions in the significant time-intervals.