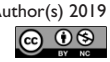
Original Article

Language *and* Speech

Language *and* Speech
1–12



DOI: 10.1177/0023830919831311
journals.sagepub.com/home/las


Non-native Listeners Benefit Less from Gestures and Visible Speech than Native Listeners During Degraded Speech Comprehension

Linda Drijvers
Centre for Language Studies, Radboud University, Nijmegen, The Netherlands;
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands;
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Asli Özyürek
Centre for Language Studies, Radboud University, Nijmegen, The Netherlands;
Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands;
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract

Native listeners benefit from both visible speech and iconic gestures to enhance degraded speech comprehension (Drijvers & Ozyürek, 2017). We tested how highly proficient non-native listeners benefit from these visual articulators compared to native listeners. We presented videos of an actress uttering a verb in clear, moderately, or severely degraded speech, while her lips were blurred, visible, or visible and accompanied by a gesture. Our results revealed that unlike native listeners, non-native listeners were less likely to benefit from the combined enhancement of visible speech and gestures, especially since the benefit from visible speech was minimal when the signal quality was not sufficient.

Keywords

non-native language processing, degraded speech, gesture, visible speech, integration, semantic integration

Introduction

As a non-native listener, understanding speech can be challenging, especially under adverse listening conditions. Previous research has shown that for native speakers, speech comprehension in adverse listening conditions can be enhanced by iconic gestures (Drijvers & Ozyürek, 2017; Holle,

Corresponding author:
Linda Drijvers, Radboud University, Centre for Language Studies, Donders Institute for Brain, Cognition and Behaviour, Wundtlaan 1, Nijmegen, 6525 XD, The Netherlands.
Email: linda.drijvers@mpi.nl

Obleser, Rueschemeyer, & Gunter, 2010; Obermeier, Dolk, & Gunter, 2011). Iconic gestures can be described as hand movements illustrating object attributes, actions, and space (e.g., McNeill, 1992). Similarly, phonological cues conveyed by visible speech, consisting of information conveyed by lip movements, tongue, and teeth, can enhance comprehension in adverse listening conditions (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollock, 1954). However, it is unknown how much non-native listeners can benefit from these visual semantic and phonological cues, or how these visual articulators interact to enhance non-native speech comprehension in clear and adverse listening conditions.

There are several reasons why the processing of information that is conveyed by visual articulators could differ in non-native listeners compared to native listeners. For example, non-native listeners might try to focus more heavily on the information that is conveyed by visual articulators to compensate for their poorer comprehension skills in the non-native language. Previous work has indeed shown that visual articulatory information conveyed by visible speech improves non-native language learning and comprehension (Hannah et al., 2017; Hazan et al., 2006; Jongman, Wang, & Kim, 2003; Kawase, Hannah, & Wang, 2014; Kim & Davis, 2014; Summerfield, 1983; Wang, Behne, & Jiang, 2008, 2009) This could especially be relevant when phoneme perception (e.g., the difference between confusable phonemes as /æ/ and /ɛ/), which is thought to be especially challenging for non-native listeners, impedes comprehension (Broersma & Cutler, 2011). Additionally, previous work indicated that both facial (i.e., visible speech) and especially gestural input (i.e., hand gestures) can help non-native tone perception, especially when phonetic demands are high, such as in noise (Hannah et al., 2017). However, most work has focused on gestures improving non-native pitch perception, tonal distinctions, and intonational patterns (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly, Bailey, & Hirata, 2017; Kelly, Hirata, Manansala, & Huang, 2014; Morett & Chang, 2015). It remains unknown how gestural semantic information enhances speech comprehension on top of visual speech in non-native listeners on a word level, and how this compares to native listeners, especially when speech is degraded.

Studies on non-native gesture processing have demonstrated that in clear speech, iconic gestures enhance non-native language comprehension and improve non-native language learning (e.g., Dahl & Ludvigsen, 2014; Hardison, 2010; Kelly, McDevitt, & Esch, 2009; Macedonia & Kriegstein, 2012; Sueyoshi & Hardison, 2005). For example, Kelly et al. (2009) have demonstrated that at initial stages, non-native listeners learn and remember novel words better when instructed with iconic gestures. In spite of this possible benefit, it is also plausible that non-native listeners might not be able to benefit from the semantic information that is conveyed by gestures on top of visible speech, especially when speech is unclear. Previous behavioral work on unimodal auditory speech processing has suggested that non-native listeners might not be able to use sentential semantic context to resolve phonemic information loss when signal quality is not clear enough. For instance, Bradlow & Alexander (2007) presented native and non-native listeners with sentences ending in low/high predictable words in plain or clear speech. Non-native listeners' comprehension of sentence-final words was only improved when both semantic and acoustic information were available (i.e., when the sentence was highly predictable and produced in clear speech), whereas native listeners could benefit from both semantic and acoustic information in combination or separately (Bradlow & Alexander, 2007; Golestani, Rosen, & Scott, 2009; Oliver, Gullberg, Hellwig, Mitterer, & Indefrey, 2012). This suggests that non-native listeners might not be able to benefit from both visual articulators when speech is unclear and auditory information is insufficient to combine the information conveyed by these two visual articulators with the auditory input.

Previous literature thus suggests that non-native listeners might utilize visual semantic and phonological cues differently than native listeners, especially in adverse listening conditions, but the contribution of these two types of visual information to speech comprehension has not yet been

studied in a joint context. A naturally following question is then whether non-native listeners can benefit from visual semantic and phonological cues in a multimodal context in a similar way to native listeners, but also whether and how these cues interact with phonological cues that are conveyed by visible speech. The aim of our study was therefore twofold: we asked whether and to what extent visible speech contributes to the enhancement of degraded speech comprehension for non-native listeners, but also whether non-native listeners experience an additive benefit from semantic information conveyed by gestures on top of enhancement of visual speech. Additionally, we aimed to compare these results with data that we have collected in previous work on native listeners (Drijvers & Ozyurek, 2017).

In Drijvers & Ozyurek (2017), we presented participants with videos of an actress uttering a clear or degraded (2- or 6-band noise-vocoding) verb while her lips were blurred, visible speech was present, or visible speech and a gesture were present. We demonstrated that native listeners benefit most from having both visible speech and an iconic gesture present as compared to having just visible speech present, or seeing the actress with her lips blurred. This effect was most prevalent at a moderate noise-vocoding level (6-band noise-vocoding) as compared to a severe noise-vocoding level (2-band noise-vocoding). These results suggested that there is a range that allows for optimal integration when the language system is weighted to an optimal reliance on auditory inputs (speech) and visual inputs (gestures and visible speech) to enhance degraded speech comprehension.

In the current study, we presented highly proficient non-native listeners of Dutch with the same stimuli and design of Drijvers & Ozyurek (2017), and compared the data from the current study to data on the native listeners reported in Drijvers & Ozyurek (2017). We focused on highly proficient participants because low proficient participants would not be able to recognize the verbs and only try to pick up information from visual cues. Investigating highly proficient listeners would thus allow us to study the enhancement that both visual articulators contribute to clear and degraded speech comprehension. We hypothesized that non-native listeners would show a similar optimal integration range around 6-band noise-vocoding due to their high proficiency, but that differences in the amount of enhancement per visual articulator might arise when they cannot effectively use visual semantic information due to their non-native listener status. For example, non-native listeners might show a similar yet diminished pattern compared to the native listeners described in Drijvers & Ozyurek (2017), because using the semantic cues that are conveyed by the gestures might be more difficult when the phonological information cannot be resolved when the speech is degraded (Bradlow & Alexander, 2007; Golestani et al., 2009; Oliver et al., 2012). Alternatively, non-native listeners might show a similar or increased use of visual articulators compared to native listeners, because their poorer comprehension of the non-native language limits them to rely on solely auditory cues.

2 Method

2.1 Participants

Twenty-three right-handed highly proficient German listeners of Dutch (7 males, MeanAge = 22.35, *SD* = 1.69) with no neurological, language, hearing, or motor disorders and normal or corrected-to-normal vision participated in this study. All participants had lived in the Netherlands for at least one year, regularly used Dutch for their studies/personal lives, and acquired Dutch after the age of 12 (meanAoA = 18.25, *SD* = 2.71) All participants partook in the LexTALE (LexTALE1) (Lemhöfer & Broersma, 2012), a five-minute non-speeded visual lexical decision test, to determine whether they indeed were highly proficient in Dutch, and only German listeners with a score

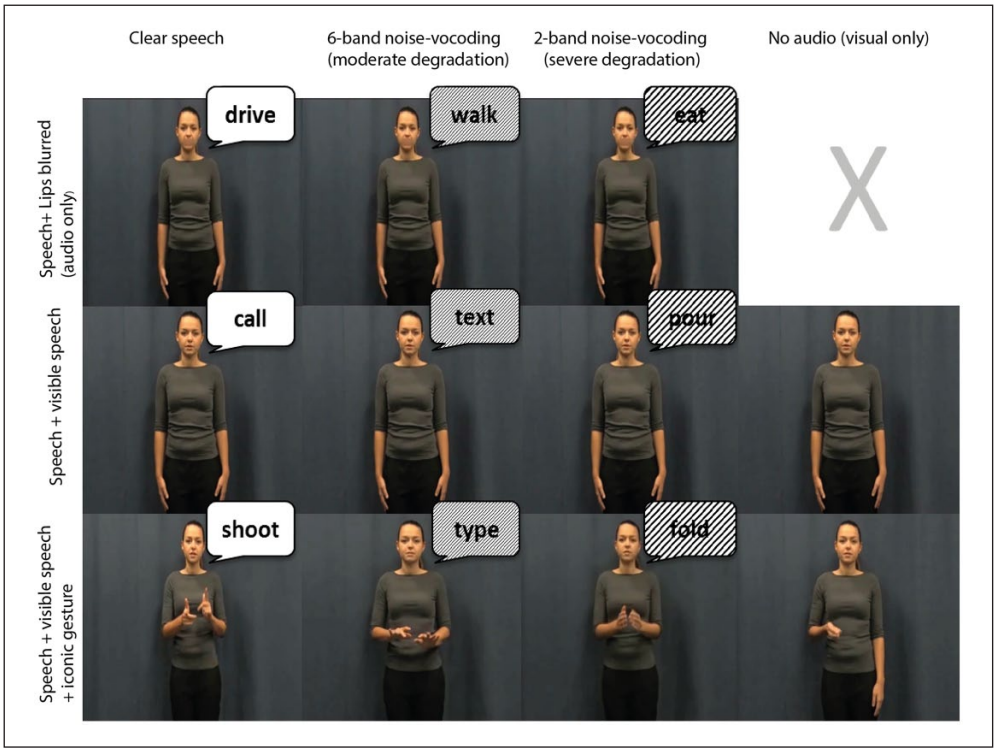


Figure 1. Overview of the design and conditions used in the experiment (picture adopted from Drijvers & Ozyurek, 2017).

above 60% were allowed to participate in the main experiment. A score of 60% and higher is thought to correlate with a B2 level or higher (upper-intermediate level). After the experiment, participants filled in an adapted version of the LexTALE test (LexTALE2) which contained the verbs that were used in the experiment, to ensure that the participants were familiar with the verbs that were used in the video (LexTALE1: MeanScore = 74.42%, *SD* = 7.97%, LexTALE2: MeanScore = 77.5%, *SD* = 11.45%). If a participant made a mistake in any of the verbs contained in the second LexTALE test, this verb was removed from the main analyses.

2.2 Stimulus Materials

The materials in this experiment are identical to the set of stimuli used in Drijvers & Ozyurek (2017) and consisted of 220 videos of a Dutch non-professional actress uttering a highly frequent Dutch action verb while she was displayed with having her lips either blurred, visible, or visible and accompanied by an iconic gesture (see Figure 1). All verbs were unique and only displayed in one condition. All gestures that were made by the actress were iconic and depicted the action verbs that she uttered (e.g., a mixing gesture resembling a whisking movement that accompanied the action verb "mixing"). The actress was asked to spontaneously make a gesture that she found representative for the verb. None of the gestures were performed in front of the face to avoid blocking the lips of the actress. The fit with the verb and iconicity of these gestures was extensively pretested and described in Drijvers & Ozyurek (2017). We only included iconic gestures that were

potentially ambiguous in the absence of speech, as this is how they are normally perceived in everyday communication (Krauss, Morrel-Samuels, & Colasante, 1991). The gestures we used had a mean recognition rate of ~59% (range: 37%–81%). We pretested these gestures by asking participants to write down which verb they associated the video with. Moreover, we asked participants to rate the gesture for how iconic that gesture was for the verb (iconicity ratings $>$ 5 on a 7-point scale (mean 6.1); for more details, see Drijvers & Ozyurek, 2017).

Every video was 2000 ms long and speech onset started on average at 680 ms after video onset. The preparation of the iconic gestures that the actress made started on average 120 ms after video onset, the stroke started on average at 550 ms, gesture retraction started at 1380 ms and gesture offset was at 1780 ms. Speech onset was on average at 680 ms, meaning that the stroke onset started on average 130 ms before speech onset, maximizing the overlap between the meaningful part of the gesture and speech for mutual enhancement and comprehension (Habets et al., 2011).

The speech in the videos was presented in clear speech, 2-band noise-vocoding, and 6-band noise-vocoding. Noise-vocoding manipulates the spectral/temporal detail of the speech while preserving the amplitude envelope of the speech signal. Noise-vocoding results in a fairly intelligible speech signal, depending on the number of bands that are used for vocoding, with more vocoding bands resulting in a more intelligible signal. For example, in 2-band noise-vocoding the signal is band-pass filtered between 50 and 8000 Hz and divided into two logarithmically spaced frequency bands, resulting in cut-off frequencies at 50, 632.5, and 8000 Hz. These frequencies were used to filter white noise in order to obtain two noise bands. The amplitude envelope of each band was extracted using half-wave rectification, multiplied with the noise bands, and recombined to form the degraded signal.

To test the different contributions of visible speech, gestures, and both articulators combined to clear and degraded speech comprehension, we divided the 220 videos over 11 conditions (20 videos per condition, identical to Drijvers & Ozyurek, 2017). In the same 3x3 design, we manipulated the number of visual articulators present in the videos (Speech + Lips blurred; Speech + VisibleSpeech; Speech + VisibleSpeech + Gesture) and the different sound levels (2-band noise-vocoding, "severe" degradation; 6-band noise-vocoding, "moderate" degradation; clear speech). Two "visual only" conditions without audio files (VisibleSpeech + Gesture; VisibleSpeech only (similar to lip-reading)) were included to test how much information participants could obtain from the visual input by itself.

2.3 Procedure

Participants were tested in a dimly lit soundproof booth, and fitted with headphones. The experimenter gave a short verbal instruction to the participant to describe the different conditions that the video could be presented in. All stimuli were presented full-screen on a 1650x1080 monitor using Presentation (Neurobehavioral systems, Inc.), at a 70 cm distance in front of the participant (identical to native listeners described in Drijvers & Ozyurek, 2017). All trials started with a fixation cross (1000 ms), followed by a free-recall task that prompted the participants to type in the verb that they perceived in the videos. After they had submitted their answer, a new trial would start after 500 ms. An answer was "correct" when the correct verb was written down, or minor spelling mistakes were made. Synonyms and category-related verbs (e.g., "to bake" for "to cook") were coded as incorrect. All participants received a different pseudo-randomization of the stimuli, and no videos were presented more than twice in a row. The stimuli were divided over blocks of 55 trials, with a self-paced rest after every block. All participants completed the experiment within 45 minutes.

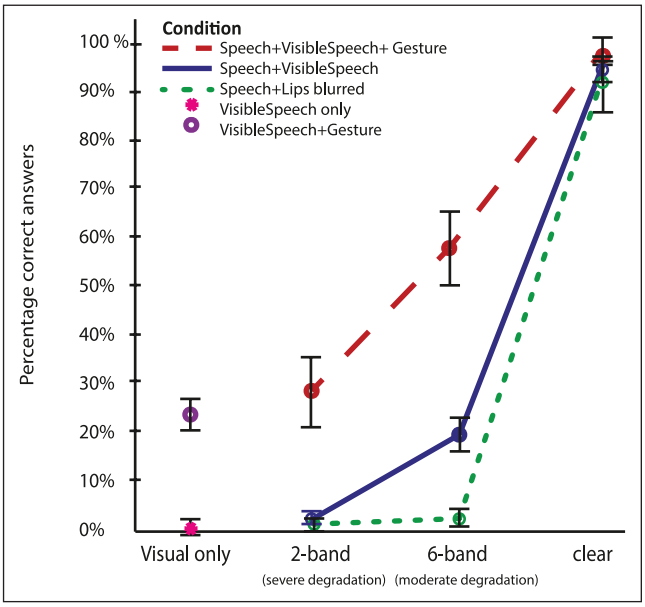


Figure 2. Percentage of correctly identified verbs (% correct) per condition.

3 Results

Following Drijvers & Ozyurek (2017), we tested whether there were differences in the number of correct answers by using a 3 (VisualArticulator: Speech+Lips blurred; Speech+VisibleSpeech; Speech+VisibleSpeech+Gesture) by 3 (Noise-Vocoding level: 2-band, 6-band, clear speech) repeated measures ANOVA. The VisualOnly conditions were tested separately. We found a significant main effect of Noise-Vocoding, $F(2,38) = 1543.23$, $p = <0.001$, partial $\eta 2 = .99$. Pairwise comparisons (Bonferroni corrected) revealed that participants significantly gave more correct answers in the clear conditions than in the 2-band noise-vocoding condition ($p = < .001$), and significantly more correct answers in the 6-band noise-vocoding condition than in the 2-band noise-vocoding condition ($p = < .001$). This thus suggests that the more noise-vocoded the signal was, the less likely it was that participants gave a correct answer.

Second, we observed a main effect of VisualArticulator, $F(1.41, 26.75) = 184{,}78$, $p = <0.001$, partial $\eta 2 = .91$, Greenhouse-Geisser corrected. Pairwise comparisons (Bonferroni corrected) revealed that participants significantly gave more correct answers in the conditions containing both visible speech and gestures than in conditions containing only visible speech ($p = < .001$), and more correct answers in conditions containing visible speech than conditions in which the lips were blocked ($p = < .001$). This thus suggests that non-native listeners were more likely to provide a correct answer when both visual articulators were present.

We observed a significant interaction between VisualArticulator and Noise-Vocoding, $F(3.06, 58.06) = 52.71$, $p = <0.001$, partial $\eta 2 = .74$, Greenhouse-Geisser corrected. Contrasts confirmed that the difference in correct answers in conditions containing visible speech and conditions containing visible speech and gestures interacted for both clear speech versus 6-band noise-vocoding, $F(1,19) = 88.45$, $p = < 0.001$ partial $\eta 2 = .83$, and 2-band noise-vocoding and 6-band noise-vocoding, $F(1,19) = 10.46$, $p = < 0.01$ partial $\eta 2 = .10$. As can be observed from Figure 2, the difference between conditions containing visible speech and gestures and conditions

containing solely visible speech seems largest in 6-band noise-vocoding. This is similar to the interaction effect that is observed in Drijvers & Ozyurek (2017) and suggests that listeners benefit most from both visible articulators at 6-band noise-vocoding.

Following previous studies (Drijvers & Ozyürek, 2017; Sumby & Pollock, 1954), we tried to confirm this effect by comparing the differences in enhancement per visual articulator as well as the enhancement per noise-vocoding level by defining three relative difference scores (A-B/100-B, i.e., enhancement types) for: a) VisibleSpeech enhancement: Speech+VisibleSpeech – Speech+Lips blurred; b) Gestural enhancement: Speech+VisibleSpeech+Gesture – Speech+VisibleSpeech; and c) Double enhancement: Speech+VisibleSpeech+Gesture – Speech+Lips blurred (see Ross et al., 2007 for a discussion of other calculation methods), divided by the maximal possible enhancement (e.g., for VisibleSpeech enhancement: 100 – Speech+Lips blurred). Note that although all enhancement types contained visible speech, we aimed to differentiate between the contribution of visible speech by itself (VisibleSpeech enhancement), gestural information occurring in the presence of visible speech (Gestural enhancement), and the contribution of both articulators combined (Double enhancement). Double enhancement could be informative about whether or not one visual articulator enhanced the enhancement of the other articulator (e.g., does the occurrence of gestural information on top of visible speech enhance comprehension on top of the enhancement of visible speech?). Moreover, these combinations more clearly mimic real-life occurrences of gesture than, for example, a condition that would contain Speech+Lips blurred+Gesture.

These difference scores were compared by means of a repeated measures ANOVA with the factors EnhancementType (VisibleSpeech, Gestural, or Double enhancement) and Noise-Vocoding (2-band, 6-band, clear speech). We observed a significant main effect of Noise-Vocoding, $F(1.35, 25{,}56) = 131.283$, $p = <0.001$, partial $\eta 2 = .87$, Greenhouse-Geisser corrected. Pairwise comparisons (Bonferroni corrected) revealed that enhancement was larger in 2-band noise-vocoding than in clear speech ($p < .001$), and larger in 6-band noise-vocoding than in 2-band noise-vocoding ($p < .001$). This thus suggests that enhancement was largest in 6-band noise-vocoding.

We observed a significant main effect of EnhancementType, $F(2, 38) = 85.93$, $p = <0.001$, partial $\eta 2 = .82$. Pairwise comparisons (Bonferroni corrected) revealed that enhancement was larger when both visual articulators were present than when one visual articulator was present ($p < .001$), and larger when both visual articulators were present than when no visual articulator was present ($p < .001$). This thus suggests that the more visual articulators were present, the more enhancement occurred.

We observed a significant interaction effect of Noise-Vocoding x EnhancementType, $F(2.71, 51.43) = 29.242$, $p = <0.001$, partial $\eta 2 = .62$, Greenhouse-Geisser corrected. Pairwise comparisons (Bonferroni corrected) revealed a significant difference between Gestural enhancement and VisibleSpeech enhancement in both 6-band noise-vocoding, $t(19) = -8.441$, $p = < 0.001$ and 2-band noise-vocoding, $t(19) = 7.644$, $p = < 0.001$, see Figure 3. There was no difference between Gestural enhancement and Double enhancement in 2-band noise-vocoding, $t(19) = -1.658$, $p = .11$, but we did observe a difference in 6-band noise-vocoding, $t(19) = -6.998$, $p < 0.001$, see Figure 3. In more detail, this thus suggests that participants benefitted most from having both visual articulators present in 6-band noise-vocoding than in 2-band noise-vocoding or clear speech, similar as was observed in Drijvers & Ozyurek (2017). However, although we observed a difference in Gestural enhancement and Double enhancement in 2-band noise-vocoding for native listeners, we did not observe this for non-native listeners.

In the VisualOnly conditions, we observed a difference between VisibleSpeech only and VisibleSpeech+Gesture, $t(19) = -15.25$, $p < 0.001$, indicating that more correct answers were given when both visible speech and gesture were present. Subsequently, we compared this Gestural enhancement (VisibleSpeech+Gesture – VisibleSpeech only/100 – VisibleSpeech only) to Gestural

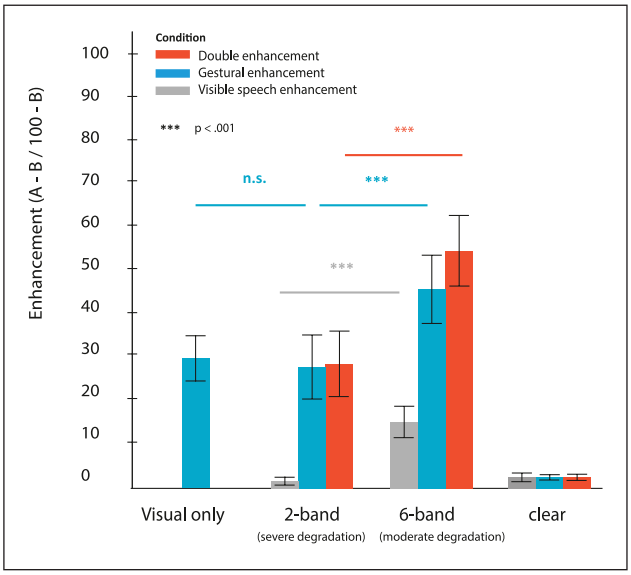


Figure 3. Enhancement effect (A-B/100-B) per visual articulator. Error bars represent *SD*, n.s. = not significant.

enhancement in 6-band noise-vocoding and 2-band noise vocoding. We only observed a difference in Gestural enhancement in the 6-band noise-vocoding condition, $t(19) = 4.683$, $p < 0.001$; but not in 2-band noise-vocoding, $t(19) = 1.566$, $p = .13$; confirming that Gestural enhancement was largest in 6-band noise-vocoding.

3.1 Comparison of Native and Non-native Listeners

Although comparisons between the data from the current dataset and data from Drijvers & Ozyurek (2017) should be made carefully, we compared the enhancement effects for VisibleSpeech enhancement, Gestural enhancement, and Double enhancement in native listeners (data from Drijvers & Ozyurek, 2017) and non-native listeners in a mixed repeated-measures ANOVA with one between-subjects factor (NativeLanguage) and two within-subject factors (EnhancementType: VisibleSpeech/Gesture/Double; Noise-Vocoding: 2-/6-band). This comparison revealed an interaction effect for EnhancementType and NativeLanguage, $F(2, 37) = 19{,}08$, $p < 0.001$, partial $\eta 2 = .508$. Contrasts confirmed that the difference in VisibleSpeech enhancement and Gestural enhancement was larger for native than non-native listeners, $F(1{,}38) = 9.77$, $p = < 0.01$ partial $\eta 2 = .21$, and that the difference in Gestural enhancement and Double enhancement was larger for native than non-native listeners, $F(1{,}38) = 4.34$, $p = 0.044$ partial $\eta 2 = .10$. Second, we observed an interaction effect for Noise-Vocoding and NativeLanguage, $F(1, 38) = 4{,}299$, $p = 0.045$, partial $\eta 2 = .102$, revealing that the difference between enhancement in 2-band and 6-band noise-vocoding was larger for native listeners than non-native listeners (see Figure 4).

4 Discussion

The first aim of this study was to investigate whether and how non-native listeners can benefit from phonological cues from visible speech and semantic cues from iconic gestures to enhance speech

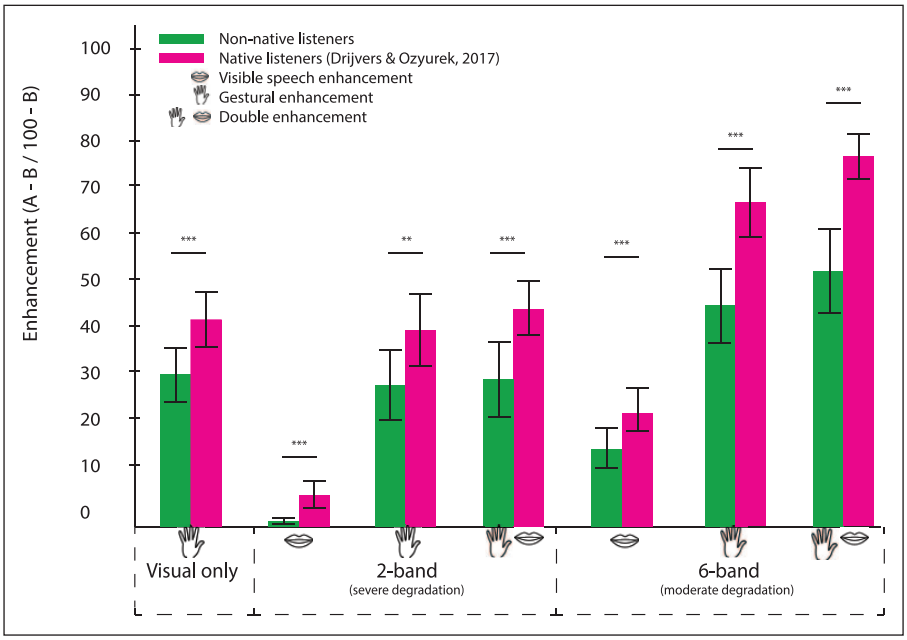


Figure 4. Enhancement effect (A-B/100-B) per language (non-native/native, adopted from Drijvers & Ozyurek (2017)). Error bars represent *SD*.

comprehension in clear and adverse listening conditions. Second, we addressed the question of how these different visual articulators interact to enhance comprehension when they are presented in a joint context, and whether there is an optimal range where there is an equal reliance on auditory inputs (e.g., speech) and visual inputs (e.g., visible speech, gestures) when non-native listeners process degraded speech. In line with previous studies (Drijvers & Ozyürek, 2017; Holle et al., 2010; Ross et al., 2007; Sumby & Pollock, 1954), we found the largest enhancement of visible speech and gestures for non-native listeners at a moderate level of noise-vocoding (6-band) as compared to severe level of noise-vocoding (2-band). However, the enhancement effects for visible speech, gesture, and both visible articulators were less pronounced for non-native listeners than the effects observed in natives in Drijvers & Ozyurek (2017).

As we observed in native listeners (Drijvers & Ozyurek, 2017), non-natives benefit most when iconic gestures were presented on top of visible speech in all noise-vocoding conditions. We did not observe differences in clear speech, which is probably due to a ceiling effect. The overall enhancement pattern for non-natives largely followed what we observed in native listeners: when both iconic gestures and visible speech were present, participants found it easiest to identify the spoken verb in the video. This enhancement benefit was larger in 6-band noise-vocoding than 2-band noise-vocoding. Note that our gestures were not completely unambiguous, which allowed speech and gesture to mutually disambiguate each other, especially in a moderately degraded context (6-band noise-vocoding). In a moderately degraded context, there are still enough auditory cues present to which to map the information that is conveyed by visual articulators, in contrast with 2-band noise-vocoding (in line with Kelly, Ozyürek, & Maris, 2010).

Interestingly, however, unlike in native listeners, we observed a difference in Gestural enhancement and Double enhancement in 6-band noise-vocoding, but not in 2-band noise-vocoding. This suggests that in 2-band noise-vocoding, the enhancement that non-native listeners experience from

iconic gestures is not aided by the presence of visible speech or the speech signal. This is most probably due to the fact that non-native listeners cannot couple the phonological cues that are conveyed by visible speech to the degraded auditory cues when noise-vocoding is too severe, which is in line with previous work on unimodal auditory speech processing (Bradlow & Alexander, 2007; Golestani, Rosen, & Scott, 2009; Oliver et al., 2012). Subsequently, non-native listeners seem to rely primarily on the semantic information of the gesture during comprehension, but they cannot make use of the extra enhancement they would get from the phonological cues that are conveyed by visible speech to benefit from double enhancement. This is corroborated by the results that were observed in the visual-only conditions. Here, native listeners were more able to correctly identify the verb than non-native listeners, suggesting that for non-native listeners it is more difficult to map both the information that is conveyed by visible speech and the information conveyed by gestures to a word that is familiar to them.

5 Conclusion

The results of the present study support our previous findings in native listeners from Drijvers & Ozyurek (2017), but also revealed some interesting differences. Whereas the enhancement from gestures was similar yet smaller for non-native listeners as compared to native listeners, the enhancement from visible speech was absent in 2-band noise-vocoding for non-native listeners. This thus indicates that in contrast to native listeners, non-native listeners might require a more intelligible signal than native listeners to benefit from visual phonological information and, subsequently, benefit from both visual articulators in a joint context. This effect was already observable at 6-band noise-vocoding, but was even larger in 2-band noise-vocoding, when even fewer phonological cues were present and signal quality was worse. This demonstrates that the optimal range for integration and reliance on auditory and visual inputs might be less liberal for non-native than native listeners because they need more phonological cues to optimally make use of the enhancement of both visible speech and gestures. This is in line with theories on speech-gesture comprehension that postulate that speech and gesture mutually enhance comprehension (Kelly et al., 2010), and explains why multimodal input, especially conveyed by visible speech and gesture, benefits non-native listeners less than native listeners.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from Netherlands Organization for Scientific Research. In loving memory of Nick Wood (†), who helped us tremendously with editing the video stimuli. We are also grateful to Gina Ginos for being the actress in the videos.

ORCID iD

Linda Drijvers https://orcid.org/0000-0001-9154-7033

References

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349.

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *Quarterly Journal of Experimental Psychology*, *64*(1), 74–95.

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, *98*(3), 813–833.

Drijvers, L., & Ozyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language & Hearing Research*, *60*, 212–222.

Golestani, N., Rosen, S., & Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Bilingualism (Cambridge, England)*, *12*(3), 385–392.

Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–1854.

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in Psychology, 8*(December), 1–15.

Hardison, D. M. (2010). Visual and auditory input in second-language speech processing. *Language Teaching, 43*(December 2009), 84.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non- native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*(3), 1740–1751.

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language and Hearing Research, 53*(April), 298–310.

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech Language and Hearing Research, 57*(December), 2090–2101.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, *49*(1), 875–884.

Jongman, A., Wang, Y., & Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *Journal of Speech Language and Hearing Research*, *46*(6), 1367.

Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *The Journal of the Acoustical Society of America*, *136*(3), 1352–1362.

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra*, *3*, 1–11.

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, *5*(July), 673. DOI: 10.3389/fpsyg.2014.00673

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*(6), 793–807.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313–334.

Kelly, S. D., Ozyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260–267.

Kim, J., & Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Computer Speech and Language*, *28*(2), 598–606.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, *61*(5), 743–754.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Macedonia, M., & von Kriegstein, K. (2012). Gestures enhance foreign language learning, *Biolinguistics*, *64*, 393–416. Retrieved from http://biolinguistics.eu/index.php/biolinguistics/article/viewFile/248/269&hl=nl&sa=X&scisig=AAGBfm0B5NzB3x1T4ypodjPFdpJcH_WQ1g&nossl=1&oi=scholarr

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: Chicago University Press.

Morett, L. M., & Chang, L. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353.

Obermeier, C., Dolk, T., & Gunter, T. C. (2011). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *CORTEX*, *48*(7), 857–870.

Oliver, G., Gullberg, M., Hellwig, F., Mitterer, H., & Indefrey, P. (2012). Acquiring L2 sentence comprehension: A longitudinal study of word monitoring in noise. *Bilingualism: Language and Cognition, 15*(May), 841–857.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699.

Sumby, W. H., & Pollock, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212.

Summerfield, A. Q. (1983). Audio-visual speech perception, lipreading and artificial stimulation. In M. E. Lutman & M. P. Haggard (Eds.), *Hearing Science and Hearing Disorders* (pp. 132–182). London: Academic Press.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716–1726.

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344–356.