At the margins of speech: orofacial expressions and acoustic cues in whispering

Marzena Żygis & Susanne Fuchs
Leibniz-ZAS, Berlin

Aside from the general consensus that gestures are related to the process of speaking (Kendon 1972, McNeill 1992), the motivation behind using gestures while speaking is still debatable. For instance, according to the *trade-off hypothesis*, if speaking becomes more difficult, the likelihood of a gesture which would 'take over' some of the communicative load is higher; and conversely, when gesturing becomes harder, speakers will rely more on speech (De Ruiter et al. 2012).

However, the hypothesis is based on voiced speech and focuses on hand gestures. It remains, however, entirely unclear what happens to orofacial expressions when (i) the speech signal becomes degenerated, and (ii) speakers do not see each other. More specifically, what happens to oral gestures if the fundamental frequency, i.e. the crucial parameter of speech, is not produced, as is the case in whispered speech? How are questions and statements realized if F0 is absent? To what extent do acoustics and orofacial expressions change if speakers whisper and do not see each other?

According to the *trade-off hypothesis*, visible orofacial motion would compensate for the F0 absence. This is in line with Dohen & Loevenbruck (2008), who showed that orofacial gestures produced while whispering decidedly enhance perception of prosodic focus in French. However, the motion would be less remarkable when speakers do not see each other.

To test the hypotheses, we conducted a motion capture experiment with 17 native speakers of German (7 male) by recording movements of their eyebrows and lip openings (see Fig. 1) in parallel to acoustic signal in four randomized blocks: (1) normal speech, visible mode; (2) normal speech, invisible mode; (3) whispered speech, visible mode; (4) whispered speech, invisible mode. In the invisible mode, the confederate and the informant were separated by an artificial wall (see Fig.2). The task of the informant was to ask a question or produce a statement in response to a sentence previously pronounced by the confederate. The sentences differed only in their final word, which was strictly controlled and consisted of a bilabial initial consonant followed by an unrounded high, mid or low vowel, e.g. *Mandel* "almond".

Several linear mixed effect models based on 2566 observations analysed the effect of speech mode [normal, whispered], visibility [visible, invisible], vowel [low, high, mid] and sentence type [question vs. statement] and their interactions on left and right eyebrow motion, lip aperture, duration and intensity of stressed syllables. Random intercepts (speaker and item) and random slopes were included as well. The results reveal that both left and right eyebrow are more raised in questions than statements (left: t=11.64; right: t=10.43) (see Fig. 3). Moreover, they are higher in invisible conditions (left: t=7.71; right: t=5.88). Both eyebrows are also more raised in whispered than normal speech (left: t=4.07; right: t=5.06). The right brow is highest in whispered invisible conditions (t=4.85). Furthermore, the lip opening is larger in (a) questions than statements (t=10.14), (b) invisible than visible conditions (t=13.63), and (c) whispered than normal speech (t=13.63). It is also highest in invisible whispered conditions. Our acoustic analysis also reveals that stressed syllables of final words were longer in questions (t=9.31), invisible conditions (t=5.39), and whispered speech mode (t= 24.85; see Fig. 4). Finally, intensity of stressed vowels being lower in whispered speech (t=-44.52) was higher in invisible conditions (t=11.71; see Fig. 4).

In summary, the results lend support to the *hand-in-hand hypothesis* as all gestures are present and even intensified in invisibility conditions. Speakers raise their eyebrows and open their mouths wider in questions even if they do not see the interlocutors, suggesting that the gestures serve speaker-internal ends. The results also point to internal compensation effects: the lack of F0 is compensated by larger lip opening and longer duration of syllables.

Fig.1. Positions of facial markers

Fig. 2. Experimental setting for an invisible mode (with an artificial wall between the speakers)

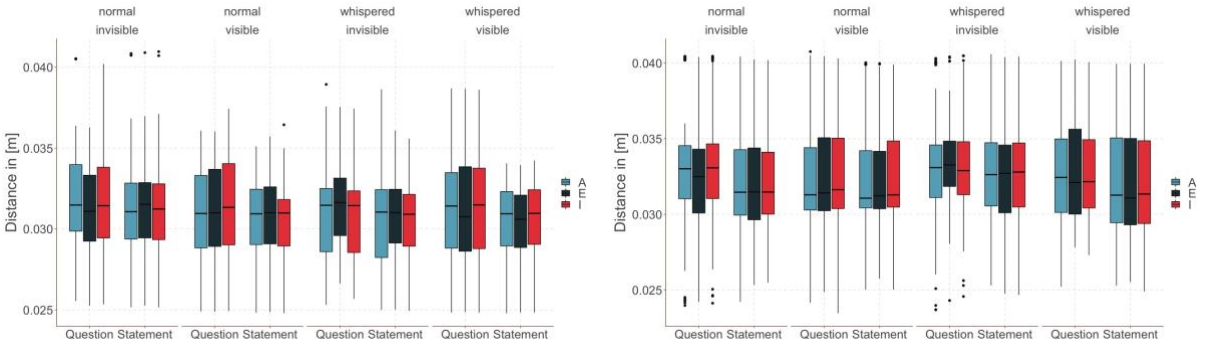

Fig. 3. 3D distance between reference marker and left eyebrow movement (left) and right eyebrow movement (right) in the sentence-final word

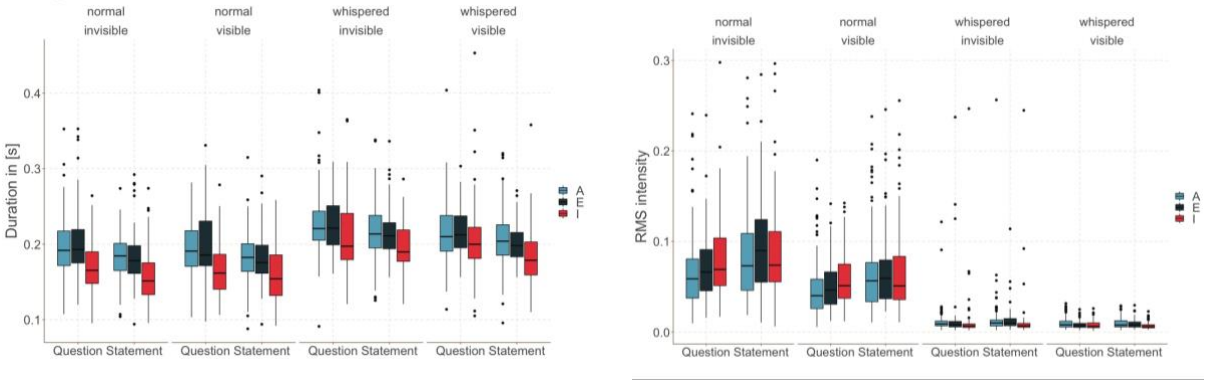

Fig. 4. Duration (left) and RMS intensity (right) of the stressed syllable in the sentence-final word

References:

De Ruiter, J. P., Bangerter, A. & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. Topics in Cognitive Science 4, 232–248.

Dohen, M., Loevenbruck, H. 2008. Audiovisual perception of prosodic contrastive focus in whispered French. Journal of the Acoustical Society of America 123, 3460-3460.

Kendon, A. 1972. Some relationships between body motion and speech. In: Sigman A. W. & B. Pope (Eds.), Studies in Dyadic Communication. New York: Pergamon Press. 177-216.

McNeill, D. 1992. Hand and Mind. Chicago: The Chicago University Press.