

Heads That Ask: Prosody of Questionhood in Whispering and Mutual (In)visibility of Speakers

Nasim Mahdinazhad Sardhaei
Humboldt-Universität / Germany
nasimmahdinazhad@gmail.com

Marzena Żygis
Leibniz-ZAS / Germany
Humboldt-Universität / Germany
zygis@leibniz-zas.de

Hamid Sharifzadeh
Unitec Institute of Technology / New Zealand
hsharifzadeh@unitec.ac.nz

Abstract

Whispering differs from voiced speech in the lack of fundamental frequency (F0) and is characterized by substantially reduced amplitude. This production study investigates how head movements are modulated in whispering when speakers are visible or not, while producing polar questions with rising intonation versus statements with flat intonation. An experiment with 15 German speakers shows that head movements are more pronounced in questions than in statements, with the strongest effects observed in normal questions under mutual visibility. The findings suggest head movements serve as a flexible multimodal prosodic resource, aligning with acoustic cues in normal speech under favorable communicative conditions while functioning independently as compensatory mechanism in whispered speech.

1 Introduction

Communication extends beyond spoken language, involving both auditory and visual channels. While speaking, we synchronize head, face, hand, and body movements with our speech, fulfilling communicative functions. These movements are closely linked to acoustic prosody cues, emphasizing key information, marking rhythmic pauses, or signaling sentence types intonation, as e.g. in questions versus statements (Cavé et al., 1996; Swerts and Krahmer, 2004; House, 2002; Dohen and Loevenbruck, 2008).

Among these various gestures, orofacial expressions and head movements such as eyebrow raises, nods, and head tilts, have been shown to carry prosodic information, complementing acoustic cues like F0. Research has consistently found significant relation between F0 and both orofacial and head movements. For example, changes in eyebrow position have been linked to variations in F0 (Cavé et al., 1996; Guaitella et al., 2009), while rigid head motions also reflect prosodic pat-

terns (Ishi et al., 2007; Yehia et al., 2002). Importantly, listeners can distinguish echoic questions from declarative statements based solely on visual head dynamics, even in simplified animations that preserved only rigid head motion indicating that head-motion cues alone can signal prosodic structure (Cvejic et al., 2010). In European Portuguese, a downward head movement (head fall) characterized yes-no questions, whereas statements displayed subtler pitch-like head motions (Cruz et al., 2015).

However, speech communication may occur in different settings, where speakers may adjust both verbal and non-verbal behavior to contextual constraints, affecting gesture production. One key factor in this interplay between speech and gesture is mutual visibility, a physical parameter that can impose varying communicative constraints, affecting speakers' strategies to convey their intended meaning effectively. Depending on whether the listener can see the speaker, the latter must adapt their verbal and non-verbal signals to the constraints of the physical medium (Bavelas et al., 2008; De Ruiter et al., 2012; Żygis and Fuchs, 2023).

The role of speakers' mutual visibility in gesture production has been subject of debate. Some studies argue that gestures primarily serve the listener, leading to fewer gestures when the listener is not visible (Bavelas et al., 1992; Krauss et al., 1995; Alibali et al., 2001; Emmorey and Casey, 2001). In contrast, others (Rimé, 1982; De Ruiter, 1998; Mol et al., 2011) propose that gestures facilitate the speaker's own speech production processes rather than serving purely communicative purposes, in which case visibility should have little effect. Notably, these studies have predominantly focused on voiced speech and specific types of gestures such as pointing. The present study manipulates speaker (in)visibility to examine how speech and head movements interact in whispered speech mode, where the lack of fundamental frequency

(F0) imposes an extra communicative constraint. In this mode, altered phonation produces voiceless acoustic signals that can challenge speech intelligibility.

Some earlier studies have examined the interaction between acoustic signals and orofacial gestures in whispered speech, with no attention to other body gestures such as head movements. For instance, [Dohen and Loevenbruck \(2008\)](#) reported that perception of focus in whispered French improved when auditory and visual cues such as eyebrow movements were combined. [Tao and Busso \(2014\)](#) showed that lip movements enhanced the accuracy of recognizing whisper. [Żygis and Fuchs \(2023\)](#) revealed higher eyebrow raising in whispered German questions. Similarly, [Mahdi-nazhad Sardhaei et al. \(2022\)](#) showed that in Farsi, facial expressions and word duration adjusted in whisper to offset the absence of F0, while [Mahdi-nazhad Sardhaei et al. \(2024\)](#) further showed that longer word durations and specific orofacial expressions acted as compensatory mechanisms in whisper to convey the intonation of a given sentence type. Building on this foundation, our study shifts the focus from orofacial expressions to head movements, investigating whether they contribute to intonational distinctions between yes–no questions and statements in whisper, where F0 is not available.

This study will examine the interplay between head movements and acoustic cues of prosody across sentence types under the above-mentioned constraints through the lens of two alternative hypotheses. The first, the trade-off hypothesis ([Bangerter, 2004](#); [Melinger and Levelt, 2004](#); [Van der Sluis and Krahmer, 2004](#); [De Ruiter, 2006](#); [Van der Sluis and Krahmer, 2007](#)), suggests that speakers tend to use gestures to ease communicative burden when verbal expression is difficult, but rely on speech when gesturing becomes demanding. The second, hand-in-hand hypothesis, proposes that gestures are produced not solely for communicative purposes but for the speaker’s own cognitive or articulatory benefits. According to this view, gestures go hand in hand with speech ([Goldin-Meadow, 2009](#); [So et al., 2009](#)). If the trade-off hypothesis holds true, we might expect head movements to compensate for the lack or distortion of F0. While, a positive relationship between head movements and acoustic parameters such as duration and mean amplitude would provide tentative support for the hand-in-hand hypothesis.

To summarize, this study aims to examine the composite (i.e., combined measures across pitch, yaw, and roll axes) velocity and magnitude of head movements, as well as their total extent of movement. Our hypotheses are:

1. Different measures of head movement will be greater in whispered questions than normally spoken questions while speakers do not see each other.
2. According to the trading hypotheses, we explore the following relations:
 - a. If the trade-off hypothesis holds true, we might observe a negative correlation between acoustic parameters (amplitude and sentence final word duration) and different head movements.
 - b. Alternatively, if the hand-in-hand hypothesis holds true, a positive correlation between acoustic parameters and head movements might emerge.

2 Methods

2.1 Participants

Fifteen native German speakers (4 male, 11 female; mean age = 33.15, SD = 12.60) with no motor, hearing, or language impairments participated. All had normal or corrected vision, provided informed consent, and received monetary compensation.

2.2 Stimuli

The study used 40 German sentences, consisting of 20 structurally identical statement–question pairs differing only in punctuation (question mark vs. period). Target words appeared in sentence-final position and were bisyllabic content words with initial syllable stress. Each initial syllable conformed to a CVC structure, beginning with a bilabial stop (/p/, /b/, or /m/) followed by unrounded vowels (/a/, /ε/, or /ɪ/). Vowels differed in height, with /a/ showing the widest lip aperture and (/ɪ/) the smallest. The target words were in sentence-final position, where intonation rises in (polar) questions and remains flat or falls in statements (for German, see [Grice et al., 2005](#)). The experiment had two contextual conditions: speech mode (normal and whisper) and visibility (visible vs. invisible). In the visible condition, participants and the confederate faced each other; in the invisible condition, an artificial wall separated them. As the result, four stimulus blocks were designed, each containing sentences classified as questions or statements:

- a. Normal speech: Speakers see each other, speakers do not see each other.

b. Whisper speech: Speakers see each other, speakers do not see each other.

To minimize order effects, both block presentation and sentence sequence within blocks were randomized for each speaker. Each speaker produced 160 sentences in total (40 sentences $\times$ 2 visibility conditions $\times$ 2 speech modes).

2.3 Data collection

The experiment was conducted in a soundproof lab with video (Canon EOS1300D, 30 Hz) and audio (Sennheiser ME62, 44,100 Hz) recordings. Participants rehearsed intonations before the session, with a confederate. In the experiment, the confederate produced either a question or a statement and the informant responded with the opposite sentence type intonation; questions followed statements and vice versa. Thus, only intonation was modified while sentence content and word order remained unchanged. For example:

- Confederates: Er mag diese Piste. (“He likes this slope.”)
- Informants: Er mag diese Piste? (“He likes this slope?”)

To ensure accurate speech mode production and proper prosody, the confederate (a trained native German speaker who remained the same across all sessions) practiced a few sentences with each participant beforehand. Throughout the experiment, this interlocutor monitored participants’ productions, and asked for repetitions when whisper quality or intonation was unsatisfactory. After data collection, a second native German speaker reviewed all utterances to verify intonational accuracy.

2.4 Data preparation

To align the timing of video and audio files of varying lengths, we used Adobe Premiere Pro 24.4 for synchronization. Audio recordings were annotated in PRAAT 6.4.20 (Boersma and Weenink, 2019), focusing on sentence-final words. Start and end times from the PRAAT TextGrid were extracted and used in a Python script to automatically segment and trim the corresponding video files. Using PRAAT 6.4.20, we manually labeled the boundaries of the sentence-final word (onset/offset), stressed syllable (onset of initial stop to syllable-final segment), and unstressed syllable (onset/offset). Our measurements included (a) sentence-final word duration and (b) amplitude difference (mean amplitude of the word-final minus word-prefinal syllable). Calculating amplitude

difference provided a normalized estimate of amplitude while partially reducing variability due to speaker–microphone distance, although it does not fully eliminate all sources of variation.

2.5 Measurement of head movement

To quantify head movement, we used OpenFace 2.0 (Baltrušaitis et al., 2018) to track facial structures. Videos of target words were processed at 29.969 FPS. OpenFace can track head movements in radians per frame. Following OpenFace conventions, we extracted head pitch (R_x , up-down nodding), yaw (R_y , left–right rotation), and roll (R_z , side-to-side tilt) to quantify head movements. Our procedures for deriving the target measures of head movements are as follows:

Velocity of head pitch, yaw, roll: The velocity of head movements represents the rate of change in head pose angles between consecutive frames, measured per unit of time. To assess head movement velocity over an entire word, we averaged the measures to summarize movement for each target word over N frames. The mean velocity over a word, considering N frames, was calculated as follows:

$$\frac{1}{N-1} \sum_{t=2}^N |\text{Velocity}(t)| \quad (1)$$

Magnitude of head pitch, yaw, roll: To this end, we defined a neutral baseline (nb) by averaging of the first 3 frames. Over an entire word, we calculated the mean absolute magnitude as:

$$\frac{1}{N-3} \sum_{t=4}^N |\text{Pitch}R_x(t) - \text{Pitch}R_x(nb)| \quad (2)$$

We also calculated composite measures which captured overall movement by combining pitch, yaw, and roll using Euclidean norms for velocity and magnitude, as follows:

Composite velocity: it gives a scalar measure of how fast the head moves overall and was calculated as:

$$\frac{1}{N} \sum_{t=1}^N \sqrt{(VR_x(t))^2 + (VR_y(t))^2 + (VR_z(t))^2} \quad (3)$$

Composite movement magnitude: it was computed per frame as the Euclidean norm of pitch,

yaw, and roll components and then averaged across N frames:

$$\frac{1}{N} \sum_{t=1}^N \sqrt{(IR_x(t))^2 + (IR_y(t))^2 + (IR_z(t))^2} \quad (4)$$

where IR_x , IR_y , and IR_z denote pitch, yaw, and roll magnitudes at frame t .

Total head movement: it was quantified as a direction-independent measure by summing movement across all axes during the word:

$$\sum_{t=2}^N \sqrt{(\Delta R_x(t))^2 + (\Delta R_y(t))^2 + (\Delta R_z(t))^2} \quad (5)$$

2.6 Statistics

We conducted statistical analyses in R Studio 4.2.3. using the lme4 (Bates et al., 2020) and emmeans packages (Lenth et al., 2018). Linear mixed models assessed the effects of Speech Mode (normal vs. whisper), Visibility (visible, invisible), Sentence Type (question, statement) on head movements. Reference levels were ‘normal speech’, ‘visible’, and ‘statement’. To investigate whether the dependent variables (Composite velocity, magnitude, total movement, as well as velocity and magnitude of head pitch, yaw, roll) exhibited the most pronounced effects under marked conditions, i.e. whispered speech, questions, and invisibility of speakers, a three-way interaction of Speech Mode, Visibility, and Sentence Type was incorporated into the models. The models accounted for random effects by including random intercepts for Word and Participant, along with by-Participant and by-Word random slopes for Speech Mode, Visibility, Sentence Type, and their interactions. Some initial models failed to converge and/or produced singularity warnings. To obtain parsimonious and identifiable random-effects structures, we followed two simplification procedures. First, random-effects terms involved in perfect correlations (i.e., correlations of ± 1) were removed, following the recommendations of (Winter, 2020). Second, when singularity persisted, we inspected the random-effects structure using the rePCA() function in the lme4 package. Following the rationale proposed by (Bates et al., 2015), the function rePCA() evaluates the effective dimensionality of the random-effects structure by performing a Principal Components

Analysis (PCA) of the estimated random-effects variance-covariance matrix. The random-effects structure is decomposed into orthogonal components, and components with very small or zero standard deviations indicate dimensions that are not supported by the data and therefore contribute to model singularity. The output also provides a rotation matrix that maps the original random-effects parameters onto the orthogonal principal components, allowing identification of redundant variance components. These components were removed through simplification of the random-effects structure (e.g., by eliminating the corresponding random slopes), resulting in a more parsimonious model. Our final model used for all dependent variables was:

```
lmer(duration ~
  Speech_mode*Visibility*sentence_type
  + (1 + Speech_mode + Visibility +
    sentence_type | participant)
  + (1 | word), data = German_data1)
```

The models were also checked with respect to (i) the normality of the residuals, (ii) their homoscedasticity and (iii) multicollinearity among the factors. To assess the significance of fixed effects in the final model, we conducted an Analysis of Variance (ANOVA) using Type III sum of squares. This approach tests each predictor’s unique contribution to the model while accounting for all other variables and their interactions. The Satterthwaite approximation was applied to estimate degrees of freedom, providing more accurate p-values in the presence of random effects. This method is particularly appropriate given the mixed-effects structure of our model and the unbalanced nature of the dataset. Finally, we then used the emmeans() function with Tukey adjustment to conduct pairwise comparisons of mean values for the dependent variables across different conditions.

Additionally, we used lmer() to examine the relationship between composite velocity and composite magnitude of head movements and two acoustic parameters of word duration and amplitude difference. We assumed that these movements may compensate for the absence of F0 in whispering. Accordingly, under the trade-off hypothesis, we would expect a negative relationship, whereas under the hand-in-hand hypothesis, we predicted a positive relationship. Continuous variables were standardized (z-scored) per speaker, and by-speaker and by-word random intercepts were included in the

models. It should be pointed out that we dispensed with a classic correlation modelling as our data included repeated (dependent) data points. An example is provided below:

```
lmer(composite_velocity ~
word_duration +
(1|participant) + (1|word))
```

3 Results

Here, we first report the results for the composite measures, and total movement, followed by a breakdown of the relative contributions of head pitch, yaw, and roll to the composite outcomes. Finally, we examine the extent to which different head movements and acoustic parameters exhibit a linear relationship.

3.1 Composite velocity

The mixed effect model revealed a significant interaction between Speech mode *Visibility*Sentence type ($F(1, 2234.02) = 9.959, p < .01$). As Figure 1 shows, head composite velocity was the greatest when speakers produced questions in the normal visible condition. Pairwise comparisons revealed greater velocity in visible normal questions than in normally produced statements under visible ($t = 3.69, p = .03$) and invisible ($t = 3.40, p = .025$) conditions.

3.2 Composite magnitude

The mixed effect model for head composite movement magnitude showed a significant interaction between Speech mode*Visibility*Sentence type ($F(1, 2217.06) = 4.125, p < .05$). Pairwise comparisons showed that head movement was greater for normally produced questions under visible condition than for statements produced in both visible ($t = -6.61, p < .001$) and invisible ($t = -5.78, p < .001$) conditions. Also, greater head magnitude was observed for visible normal questions compared to invisible whispered statements ($t = -3.53, p = 0.046$, Figure 1).

3.3 Total movement

For total head movement, the interaction between Speech mode*Sentence type turned out to be significant ($F(1, 2231.89) = 26.705, p < .001$). As Figure 2 (right) shows, normally produced questions had the largest total movement. Pairwise comparisons indicated that normally produced questions had a larger total movement than normal statements ($t = -6.23,$

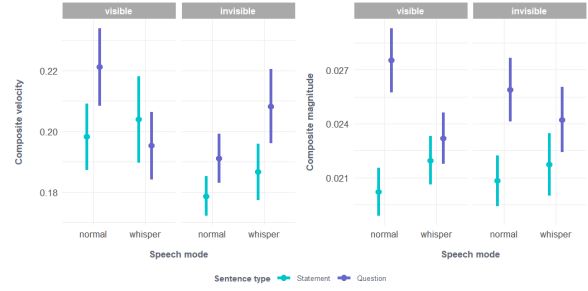


Figure 1: Composite velocity (left) and magnitude (right) of head movement in the sentence-final word: interaction between speech mode, (in)visibility, and sentence type.

$p < .001$). on the other hand, whispered questions showed a higher total movement than whispered ($t = -3.12, p = 0.01$) and normal ($t = -3.02, p = 0.03$) statements.

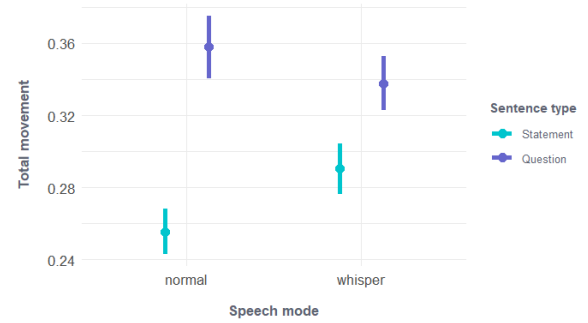


Figure 2: Total head movement showing an interaction between speech mode, and sentence type.

3.4 Contribution of individual axes to composite head movement

To clarify which axis of head movement primarily drove the composite results, we decomposed the composite measures into their individual pitch (Rx), yaw (Ry), and roll (Rz) components. Among these, head pitch emerged as the dominant contributor. Both pitch velocity and magnitude yielded significant interactions between Speech mode*Visibility*Sentence type (velocity: $F(1, 2233.54) = 10.916, p < .001$; and magnitude: $F(1, 2201.21) = 4.629, p < .05$; see Figures 3 and 4). Specifically, the head pitch velocity was the greatest when speakers produced questions in selected conditions (visible normal and whisper invisible). However, pairwise conditions revealed only few significant differences with greater pitch velocity in normal questions under visible condition than normally produced statements under visible ($t = 3.74, p = 0.013$) and invisible ($t = 3.96, p = 0.017$) con-

ditions. The close correspondence between pitch and composite patterns suggests that head nodding accounts for the largest share of variance in overall head displacement.

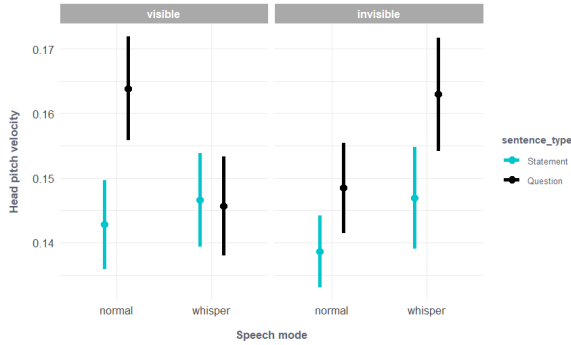


Figure 3: Head pitch velocity in the sentence-final word: three-way interaction between speech.

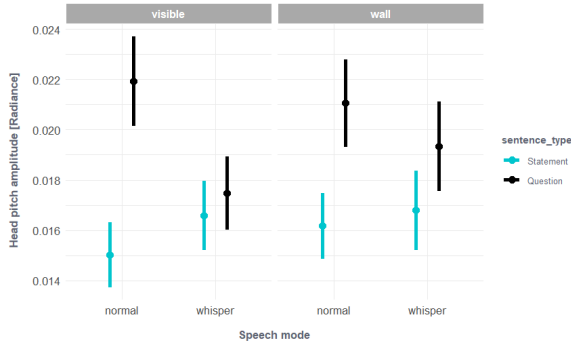


Figure 4: Head pitch magnitude in the sentence-final word: three-way interaction between speech mode, visibility, and sentence type.

By contrast, head roll movement did not show any significant effects. Only, yaw magnitude yielded a significant two-way interaction between speech mode and sentence type ($F(1, 2213.35) = 4.914, p < .05$), with questions produced with greater yaw, and a larger difference between questions and statements in normal speech than in whisper ($t = 2.87, p = 0.03$).

3.5 Relationship between acoustic cues and head movements

At this stage, we analyzed the relationship between head movements (composite velocity and composite magnitude) and two acoustic parameters including word duration and mean amplitude difference. All continuous variables were z-scored for each speaker. Although the observed relationships were small, some were statistically significant.

The analysis revealed a negative relationship be-

tween word duration and head composite velocity in both normal ($t = -3.22, p < .001$) and whisper ($t = -2.53, p = 0.011$) modes, indicating faster movement for shorter words (see Figure 5). In contrast, composite head movement magnitude was positively correlated to word duration only in whispered speech ($t = 2.82, p < .01$), with longer words showing larger movements.

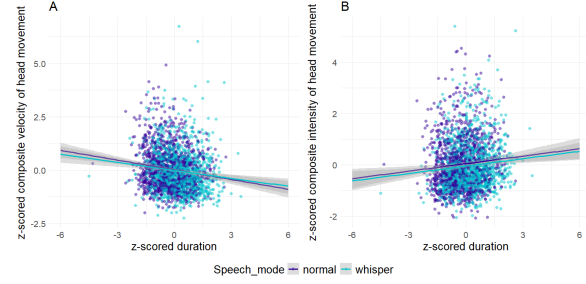


Figure 5: Scatterplots for sentence-final word duration in relation to head movement composite velocity and magnitude.

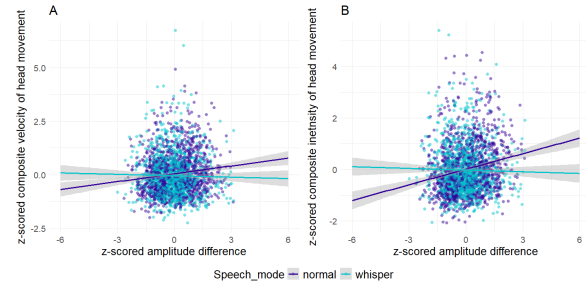


Figure 6: Scatterplot for amplitude difference in relation to head movement composite velocity and magnitude.

Mean amplitude difference and head composite velocity were positively correlated in normal speech mode ($t = 4.70, p < .0001$, see Figure 6) suggesting louder words were associated with faster head movement. No significant result was observed for whispered speech. Similar results were found for composite head movement magnitude, with a positive relationship to mean amplitude difference only in normal speech ($t = 5.23, p < .0001$).

4 Discussion

To our knowledge, this is the first study to examine head movements in conveying sentence-type intonation under communicative constraints, particularly in whispered speech, while manipulating interlocutor mutual visibility. The findings are discussed in relation to the trade-off and hand-in-hand hypotheses and the broader literature on multimodal prosody. We focused on sentence-final

words where intonational contrasts between polar questions and statements are primarily realized. Even within this brief part of the sentence, significant differences emerged across all head movement measures with faster and larger head movements in questions than in statements. This pattern was evident in all head movement measures. Consistent with earlier research (e.g. Cvejic et al., 2010; Yehia et al., 2002), these results confirm that rigid head motion carries prosodic information and contribute to the signaling of sentence type. Questions, typically characterized by rising intonation, appear to involve enhanced multimodal prosodic marking compared to statements. The ability of listeners to distinguish sentence types based solely on head motion (Cvejic et al., 2010) further supports the communicative relevance of the patterns observed in our study.

Our initial prediction was that the strongest head gestures would occur under the most communicatively constrained condition; that is, whispered questions produced in the absence of speakers' mutual visibility. However, the results did not confirm this hypothesis. Instead, the most pronounced head movements were consistently found in the least constrained condition, i.e., normally produced questions in the visible condition with the largest effects across composite velocity, composite magnitude, and total movement. Some previous studies have also reported increased gesture production when speaker and listener can see each other than when there is a visual barrier between the two (e.g. Alibali et al., 2001; Mol et al., 2011; Barker and Kita, 2026). Boosting of head movements in normally produced questions when interlocutors are visible may reflect a general prosodic marking strategy that is amplified when full communicative resources are available. Under these circumstances, head movements appear to function in parallel to acoustic prosodic cues such as F0, reinforcing the sentence type contrast through visual signaling. This redundancy may reflect an audience-oriented strategy (Bavelas et al., 1992), whereby speakers enhance communicative clarity when resources are optimal.

The role of visibility, however, proved selective rather than uniformly amplifying. Although a significant interaction between speech mode, visibility, and sentence type indicates that visibility contributes to gestural behavior, this effect is not consistent across all measures as total head movement was unaffected by (in)visibility. Thus, visibility

may not globally influence head gesture production but selectively modulate specific kinematic parameters.

To better interpret the findings, we examined the relative contributions of individual head movement axes revealing head pitch as the primary vehicle of prosodic head movement in our data. Head pitch emerged as the most prominent marker of sentence type contrasts, showing increased velocity and magnitude for normally produced questions, when the interlocutor was visible while head yaw and roll playing very small roles. This close correspondence between head pitch patterns and the composite measure indicated that head nodding accounts for the largest share of variance in overall head displacement, and that the sentence-type contrasts observed in the composite measures are primarily driven by up-and-down head rotation. Another point to notice was significant magnitude of head yaw movement and its contribution to visually marking sentence-type distinctions. The co-occurrence of pitch and yaw magnitude in questions suggests that these two axes may not operate entirely independently. Instead, speakers may use coordinated gesture combining up-down (pitch) and left-right (yaw) head motions. In other words, a speaker may nod while slightly turning the head. The resulting movement is not purely nodding or turning, but a smooth trajectory that has incorporates both components. Such multi-axis coordination in prosodic gesturing has been noted in previous accounts of head movement behavior (Ishi et al., 2007), (Yehia et al., 2002), and may reflect a naturally integrated gestural strategy for marking sentence-type boundaries visually.

Our results can be interpreted considering two hypotheses: the trade-off hypothesis which posits that gestures are produced to support listener comprehension (Alibali et al., 2001), and the hand-in-hand hypothesis according to which gesturing helps the speaker reduce cognitive load (Goldin-Meadow, 2009).

In this study, the relationship between head movements and acoustic parameters (word duration and mean amplitude difference) does not straightforwardly support either hypothesis alone. Instead, it points to a flexible system in which head movements reflect different mechanisms depending on speech mode, sentence type, and speaker visibility, with patterns consistent with both accounts.

For sentence-final word duration, shorter words were accompanied by faster composite velocity in

both normal and whispered speech which suggests a timing-based coordination mechanism: head movement dynamics are adjusted to align with the temporal structure of speech rather than reflecting a simple scaling predicted by either hypothesis. On the other hand, in whispered speech, longer words were associated with greater composite magnitude which may be consistent with the hand-in-hand hypothesis. However, rather than drawing a definitive conclusion, we interpret this parallel enhancement of acoustic information and head movements as a compensatory strategy for the absence of F0 and degraded amplitude where speakers may rely on alternative prosodic cues such as longer duration coupled with more pronounced head movements to convey intended intonation. Together, these patterns indicate that in whispered speech, speakers flexibly adjust their head movements in tandem with word duration balancing temporal constraints with the need to compensate for reduced acoustic prosodic cues.

The amplitude difference analysis revealed louder sentence-final words were associated with faster and larger head movements. The synchronous strengthening of acoustic and gestural signals supports the hand-in-hand hypothesis (Goldin-Meadow, 2009) viewing gesture and speech as parallel systems driven by the same underlying demands. In normal speech, speakers may coordinate acoustic and visual cues of prosody either to enhance listeners' perception of prosodic contrasts or increase communicative effectiveness of their own utterance (Van der Sluis, 2005; Van der Sluis and Krahmer, 2007). Moreover, this finding resonates with the motor-prosodic coupling hypothesis (McNeill, 1992), which holds that prosodic prominence engages greater biomechanical energy across both speech articulators and the body. Louder speech requiring greater articulatory effort may co-occur with enhanced head and upper-body movement activity (Hadar et al., 1983; Munhall et al., 2004). In this view, greater movement magnitude reflects a coordinated motor response to prosodic emphasis rather than a direct consequence of acoustic intensity alone. No such relationship was found in whispered speech, likely because amplitude differences are much degraded in this speech mode.

5 Conclusion

In summary, by examining the relationship between the acoustic signal and oro-facial expressions when

speakers (i) speak normally or whisper, and (ii) do or do not see each other, we found the overall advantage of visible, normally spoken questions over all other conditions. The results are interpreted in terms of two hypotheses, i.e., the hand-in-hand and trade-off hypotheses. The relationship between acoustic properties and head gestures proved to be mode-dependent rather than uniform. In normal speech, the positive coupling between amplitude difference and both composite velocity and magnitude points to a hand-in-hand mechanism, whereby acoustic and head gesture are co-produced and mutually reinforcing. In whispered speech, however, this coupling breaks down, and head movements appear to operate more independently: longer words are accompanied by greater head movement magnitude, while shorter words trigger faster gestural velocity suggesting a compensatory mechanism to preserve prosodic contrasts in the absence of F0, while also balancing temporal constraints by faster movement. Although head movements may require more effort than other orofacial expressions, the observed patterns indicate that speakers use their heads as an important visual resource, reinforcing prosodic marking across different communicative conditions. Nevertheless, previous research has also shown that orofacial movements serve as important cues for marking sentence-type intonation in whisper. Further perceptual studies could investigate the relative contributions of head movements and orofacial expressions in distinguishing different prosodic patterns in whisper vs. normal speech. In particular, such studies could determine whether head movements play a compensatory role in whispering by helping listeners recover prosodic distinctions when F0 is absent. The present analyses focused on direction-independent composite measures capturing how large and how fast speakers moved their heads to mark prosodic contrasts, without distinguishing the specific orientations of those movements. While this approach provides a summary of prosodic head gesturing, future studies incorporating head movement trajectories would provide a more detailed account of the directional and temporal dynamics of prosodic head gesturing. Such work would benefit further from being conducted in naturalistic communicative settings with spontaneous interactions to capture a more ecologically valid picture of how head movements signal sentence-type intonation in everyday, particularly whispered speech. Our findings are based on a relatively small sample of 15 participants, which

may constrain the generalizability of the results. This limitation highlights the need for further research involving larger and more diverse participant samples. Finally, it remains an open question whether the patterns observed in German generalize to other languages with different prosodic systems, or to sentence types beyond the polar question–statement distinction examined here.

6 Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Additionally, we extend our gratitude Jorg Dreyer for his support in running the experiment, and to our participants for their participation. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, project “Audio-visual prosody of (semi-)whispered speech” ZY 117/4-1).

References

- M. W. Alibali, D. C. Heath, and H. J. Myers. 2001. Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44(2):169–188.
- T. Baltrušaitis, C. Ahuja, and L. P. Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
- A. Bangerter. 2004. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419.
- J. Barker and S. Kita. 2026. The addressee’s face as a cue for gesture production. *Language, Cognition and Neuroscience*, pages 1–18.
- D. Bates, R. Kliegl, S. Vasishth, and H. Baayen. 2015. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- D. Bates, M. Maechler, B. Bolker, and S. Walker. 2020. Package ‘lme4’ (version 1.1-26) [computer software]. Comprehensive R Archive Network.
- J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2):495–520.
- J. B. Bavelas, N. Chovil, D. A. Lawrie, and A. Wade. 1992. Interactive gestures. *Discourse Processes*, 15(4):469–489.
- P. Boersma and D. Weenink. 2019. Praat: Doing phonetics by computer, version 6.0.55. <https://www.praat.org>.
- C. Cavé, I. Guaïtella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. 1996. About the relationship between eyebrow movements and f0 variations. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, volume 4, pages 2175–2178.
- M. Cruz, M. Swerts, and S. Frota. 2015. Variation in tone and gesture within language. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, pages 1–5.
- E. Cvejic, J. Kim, C. Davis, and G. Gibert. 2010. Prosody for the eyes: Quantifying visual prosody using guided principal component analysis. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1433–1436.
- J. P. De Ruiter. 1998. *Gesture and speech production*. Ph.D. thesis, Radboud University Nijmegen, Nijmegen.
- J. P. De Ruiter. 2006. Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, 8(2):124–127.
- J. P. De Ruiter, A. Bangerter, and P. Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2):232–248.
- M. Dohen and H. Loevenbruck. 2008. Audiovisual perception of prosodic contrastive focus in whispered french. *Journal of the Acoustical Society of America*, 123(5):3460.
- K. Emmorey and S. Casey. 2001. Gesture, thought and spatial language. *Gesture*, 1(1):35–50.
- S. Goldin-Meadow. 2009. How gesture promotes learning throughout childhood. *Child Development Perspectives*, 3(2):106–111.
- M. Grice, S. Baumann, and R. Benzmüller. 2005. German intonation in autosegmental-metrical phonology. In S.-A. Jun, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, pages 55–83. Oxford University Press, Oxford, UK.
- I. Guaïtella, S. Santi, B. Lagrue, and C. Cavé. 2009. Are eyebrow movements linked to voice variations and turn-taking in dialogue? an experimental investigation. *Language and Speech*, 52(2-3):207–222.
- U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose. 1983. Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129.
- D. House. 2002. Intonational and visual cues in the perception of interrogative mode in swedish. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP/Interspeech)*, pages 1957–1960.

- C. T. Ishi, H. Ishiguro, and N. Hagita. 2007. Analysis of head motions and speech in spoken dialogue. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 670–673.
- R. M. Krauss, R. A. Dushay, Y. Chen, and F. Rauscher. 1995. The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology*, 31(6):533–552.
- R. Lenth, H. Singmann, J. Love, P. Buerkner, and M. Herve. 2018. Package ‘emmeans’ (version 4.0-3) [computer software]. Comprehensive R Archive Network. <http://cran.r-project.org/package=emmeans>.
- N. Mahdinazhad Sardhaei, M. Żygis, and H. Sharifzadeh. 2022. How do our eyebrows respond to masks and whispering? the case of persians. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2023–2027.
- N. Mahdinazhad Sardhaei, M. Żygis, and H. Sharifzadeh. 2024. Facial expressions in different communication settings: A case of whispering and speaking with a face mask in farsi. *Language and Cognition*, 16(4):1639–1673.
- D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.
- A. Melinger and W. J. Levelt. 2004. Gesture and the communicative intention of the speaker. *Gesture*, 4(2):119–141.
- L. Mol, E. Krahmer, A. Maes, and M. Swerts. 2011. Seeing and being seen: The effects on gesture production. *Journal of Computer-Mediated Communication*, 17(1):77–100.
- K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137.
- B. Rimé. 1982. The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology*, 12(2):113–129.
- W. C. So, S. Kita, and S. Goldin-Meadow. 2009. Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33(1):115–125.
- M. Swerts and E. J. Krahmer. 2004. Congruent and incongruent audiovisual cues to prominence. In *Proceedings of the 2nd International Conference on Speech Prosody*, pages 69–72.
- F. Tao and C. Busso. 2014. Lipreading approach for isolated digits recognition under whisper and neutral speech. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1154–1158.
- I. Van der Sluis. 2005. *Multimodal reference*. Ph.D. thesis, Katholieke Universiteit Brabant, Tilburg. Unpublished Doctoral Dissertation.
- I. Van der Sluis and E. Krahmer. 2007. Generating multimodal references. *Discourse Processes*, 44(3):145–174.
- I. F. Van der Sluis and E. J. Krahmer. 2004. The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, volume 4.
- B. Winter. 2020. *Statistics for Linguists: An Introduction Using R*. Routledge.
- H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. 2002. Linking facial animation, head motion, and speech acoustics. *Journal of Phonetics*, 30(3):555–568.
- M. Żygis and S. Fuchs. 2023. Communicative constraints affect oro-facial gestures and acoustics: Whispered vs normal speech. *Journal of the Acoustical Society of America*, 153(1):613–626.