

ARTICLE

Facial expressions in different communication settings: A case of whispering and speaking with a face mask in Farsi

Nasim Mahdinazhad Sardhaei¹, Marzena Żygis^{1,2} and Hamid Sharifzadeh³

¹Leibniz-Zentrum für Allgemeine Sprachwissenschaft, Berlin, Germany; ²Humboldt Universität, Berlin, Germany and ³Unitec Institute of Technology, Auckland, New Zealand

Corresponding author: Nasim Mahdinazhad Sardhaei; Email: sardhaei@leibniz-zas.de

(Received 14 December 2022; Revised 12 February 2024; Accepted 23 February 2024)

Abstract

This study addresses the importance of orofacial gestures and acoustic cues to execute prosodic patterns under different communicative settings in Farsi. Given that Farsi lacks morpho-syntactic markers for polar questions, we aim to determine whether specific facial movements accompany the prosodic correlates of questionhood in Farsi under conditions of degraded information, that is, whispering and wearing face masks. We hypothesise speakers will employ the most pronounced facial expressions when whispering questions with a face mask to compensate for the absence of F0, reduced intensity and lower face invisibility. To this end, we conducted an experiment with 10 Persian speakers producing 10 pairs of statements and questions in normal and whispered speech modes with and without face masks. Our results provide support to our hypotheses that speakers will intensify their orofacial expressions when confronted with marked conditions. We interpreted our results in terms of the ‘*hand in hand*’ and ‘*trade-off*’ hypothesis. In whispered speech, the parallel realisation of longer word duration and orofacial expressions may be a compensatory mechanism for the limited options to convey intonation. Also, the lower face coverage is mutually compensated for by word duration and intensified upper facial expressions, all of which in turn support the trade-off hypothesis.

Keywords: face mask; multimodal prosody; orofacial gestures; sentence type; trading relations; whispering

1. Introduction

1.1. *The multimodality of speech prosody*

Recent research has shown that human communication is essentially a multimodal form of signalling in which vocal and visual modes of communication form an integrated system of meaning (e.g., Holler & Levinson, 2019; Kendon, 2004; Levinson & Holler, 2014; McNeill, 1992). Indeed, communication partners mutually profit



from a large repertoire of visual cues derived from physical movements, which combine with verbal signals to fulfil their communicative goals. One of the primary questions on the association between visual and verbal components of language is whether – and if so, how – various gestural movements accompanying speech contribute to prosody, an inherent part of human communication.

It has been long believed that prosody is manifested through tonal, temporal and spectral properties of speech, with several studies almost exclusively viewing prosody via purely auditory/acoustic channels. However, more recently, there has been a growing awareness that the production and perception of prosody are multimodal and that facial expressions as well as other forms of body gestures could serve similar communicative functions as the auditory cues of prosody (e.g., Guellai et al., 2014; Krahmer & Swerts, 2007; McNeill et al., 2001; Mendoza-Denton & Jannedy, 2011). In face-to-face scenarios, speakers naturally use both voice and body movements to communicate their intentions as they can both hear and see their interlocutors. It is thus reasonable to suppose that prosody is expressed via visual modality in addition to the acoustic properties of speech.

In this regard, human body movements have attracted a great deal of research demonstrating the contribution of various types of visual cues to the auditory properties of prosody (e.g., Ambrazaitis & House, 2017; Dohen et al., 2004; Esteve-Gibert & Prieto, 2013; Guellai et al., 2014; Holler & Levinson, 2019; Kendon, 2004; Krahmer & Swerts, 2007; Prieto et al., 2015; Shattuck-Hufnagel & Ren, 2018). However, research on the role of facial movements has primarily focussed on the emotional correlates of speakers' utterances (Bould & Morris, 2008; Delis et al., 2016; Sato & Yoshikawa, 2004; Schmidt et al., 2009; Sowden et al., 2021), while the number of studies investigating how various facial movements contribute to the prosody of spoken language in non-emotional contexts is significantly smaller (see Section 1.2 for details).

This study examines the relationship between orofacial gestures and the acoustic cues of intonation in polar questions as opposed to statements. We aim to find out if this relationship changes in different communicative settings – specifically those of whispered versus voiced speech modes and with or without a protective face mask. In section 1.2., the results of previous empirical studies on the role of orofacial gestures in marking the intonation of sentence will be explained. In Section 1.3, we discuss existing hypotheses on the relationship between speech and gesture. Next, whispered speech and prosodic correlates in this mode of speech will be introduced. In Section 1.5, we report briefly on the existing literature on the effect of wearing protective face masks on speech perception and production. In Sections 2 and 3, we present our experiment and report its results. Finally, in Section 4 we discuss the results and provide conclusions.

1.2. The role of facial movements and auditory cues of prosody in signalling sentence type

To date, studies concentrating on the integration of facial cues and auditory information for signalling a coherent percept have confirmed temporal correlations between various facial expressions and the acoustic properties of prosody. For instance, it has been regularly reported that changes in fundamental frequency and amplitude, as the acoustic correlates of prosody are associated with simultaneous

eyebrow movements (Cavé et al., 1996; Yehia et al., 2002). Such movements and other articulatory gestures have accordingly been mapped onto the intonation of sentence types, including questioning with utterance-final rising pitch and responding with falling intonation (Bavelas et al., 2014; Borràs-Comes & Prieto, 2011; Cruz et al., 2017; Hömke et al., 2022; House, 2002; Miranda et al., 2021; Nota et al., 2021).

Based on audiovisual prosody literature, both the use of facial gestures and the movement patterns of eyebrows may vary depending on the pragmatic function and type of question (Cruz et al., 2015, 2017). For instance, eyebrow raising in spoken Dutch marks a request for repair (e.g., *What?; Huh?*) or clarification (e.g., *Johnny Smith?*), while eyebrow furrowing marks a restricted request for repair (e.g., *He did what?*; see Hömke, 2019; Hömke et al., 2022). Echo questions in American English have been found to be differentiated by eyebrow raising (Srinivasan & Massaro, 2003), whereas the expression of echo questions in Catalan is associated with lowered eyebrows (Borràs-Comes & Prieto, 2011). Comparing polar questions in Dutch and Catalan, Borràs-Comes et al. (2014) reported that eyebrow raising and eye gazing were present in both languages. Similarly, the production of polar questions in European Portuguese is accompanied by head movements along with eyebrow raising (Cruz et al., 2015). Raised eyebrows are also present in polar questions and statements in French (Torreira & Valtersson, 2015). Conversely, Swedish speakers use eyebrow lowering as a visual cue for conveying polar questions and presenting statements with eye narrowing (House, 2002). Lowered eyebrows are also employed to mark the production of *wh*-questions in Brazilian Portuguese (Miranda et al., 2019). In addition, a comparison of Brazilian Portuguese and Mexican Spanish (Miranda et al., 2020) showed that while assertions are produced with different visual cues in the two languages, both echo and polar questions are produced by lowering the eyebrows, tightening the eyelids and wrinkling the nose.

Visual prosodic marking of an utterance as a question varies not only across sentence types and pragmatic functions, but also across various languages or different varieties of the same language. The facial expressions accompanying questions, such as eyebrow movements, may differ depending on the linguistic question-marking strategies of the language concerned in terms of lexical input or specific syntactic structures. In many languages, questions are indicated by question particles (e.g., French *est-ce-que*, English *wh*-words) or specific word order, typically an inversion of the subject and verb positions (e.g., Dutch *Heeft hij het boek gelezen?* 'Has he read the book?'; see Borràs-Comes et al., 2014). On the other hand, some languages such as Catalan and Brazilian Portuguese, mark questionhood through a change in intonation. From both production and perception perspectives, cross-linguistic research comparing the audiovisual correlates of different sentence types has reported more or less language-specific differences in the use of visual cues to mark or perceive sentences. Some of these studies have suggested a functional trade-off distribution between visual signals and lexico-syntactic strategies encoding questionhood. In languages that exploit morpho-syntactic question-marking strategies, visual features have been shown to have a weaker cue strength than auditory ones for the distinction between statements and questions. This was observed, for instance, with Swedish (House, 2002) and English (Srinivasan & Massaro, 2003). By contrast, when a language system lacks morpho-syntactic question-marking strategies, the visual modality can act in a trade-off compensatory fashion to express questionhood. Also, in languages or across language varieties, for cases where the auditory/tonal features are not very informative, speakers may be more sensitive to visual cues (Cruz et al.,

2017). For instance, earlier research on a few languages such as Catalan and Dutch (Borràs-Comes et al., 2014; Borràs-Comes & Prieto, 2011; Crespo-Sendra et al., 2013) indicates that in a language (Catalan) that uses same intonational contour but a different pitch range for two distinct pragmatic meanings, visual cues can be more relevant. Conversely, in Dutch, which uses distinct intonational contours, visual cues may play a secondary role. In a cross-linguistic study, Crespo-Sendra et al. (2013) demonstrated that in perceptual evaluations, Catalan listeners give more weight to facial cues, including eyebrow furrowing, eyelid closing and forward head-tilting to perceive neutral and focussed polar questions, whereas Dutch listeners rely more on prosodic cues. Another comparative analysis of Dutch and Catalan (Borràs-Comes et al., 2014) found that although both languages showed similar distributions of eyebrow raising in questions, Catalan speakers used more visual cues (eyebrow raising and eye gazing) than Dutch ones. However, a perceptual study testing the role of facial gestures in the absence of a tonal contrast across two language varieties of European Portuguese (Cruz et al., 2017) suggested that visual cues alone are not sufficient for sentence-type identification, with listeners relying more on auditory information than the visual information.

Taking the two points above into account, that is the function of questioning and language-specific differences in question-marking strategies, this study aims to investigate the role of intonation and facial movements in the production of two sentence types, namely declaratives and polar questions in Farsi, the official language of Iran. In Farsi, the syntactic structure of polar questions and declaratives is generally identical as this language does not rely on a specific morpho-syntactic strategy for the production of polar questions. Two major distinguishing factors characterising polar questions in Farsi are: (i) greater pitch excursion as well as final lengthening on the last accentual phrase of polar questions (Sadat-Tehrani, 2011), and (ii) intonational phrase boundary tone: while declaratives use an L-L% boundary tone, polar questions are marked by an L-H% one (Sadat-Tehrani, 2011). The act of intonational change to mark questionhood in Farsi is a feature it has in common with many other languages such as European Portuguese (Frota, 2002), Mandarin Chinese (Zeng et al., 2004), Spanish and Italian (Arvaniti et al., 2006). Nevertheless, Farsi is one of the unexplored languages in the field of multimodality and audiovisual speech research, and we are unaware of any studies investigating both acoustic cues and orofacial expressions in the linguistic context of this language. To this end, we aim to find out whether there are any specific patterns of facial movements accompanying the prosodic correlates of questionhood in Farsi, and if so, how they change in different communicative settings. Given the fact that Farsi does not employ morpho-syntactic markers to express polar questions and comparing it with the set of other languages described earlier in the literature, we would expect similarities in the relationship between visual signals and acoustic cues marking sentence types for Farsi and that set of languages in which polar questions are especially cued by prosody.

1.3. *The relationship between speech and orofacial gestures*

Despite much research, the exact nature of gestural-acoustic interplay is still far from fully understood, and the motivation behind using gestures in speech is still under scrutiny. While some studies, on the one hand, have declared that speakers produce gestures to facilitate the perception of their message by listeners (Alibali et al., 2001),

it has also been proposed, on the other hand, that gesturing helps the speaker to reduce cognitive load (Goldin-Meadow et al., 2001). Closely related to these suggestions are two hypotheses. One possibility is that there is a *trade-off* relation between gesture and speech in terms of the communicative load (Bangerter, 2004; De Ruiter, 2006; Melinger & Levelt, 2004; Van der Sluis & Krahmer, 2007). According to this *trade-off* hypothesis, there is a two-way compensatory relationship between gesture and speech, in which the difficulty of each modality increases the likelihood of intensification of another modality to take over some of the communicative burden. When speaking becomes difficult, speakers will resort to gesturing to compensate for the reduction in speech modality; and conversely, when gesturing becomes difficult, they will rely more on speech. An alternative conjecture is the *hand-in-hand* hypothesis, which views the relationship between gestures and speech as parallel or redundant rather than compensatory in the sense that gestures basically express information that can be derived from the spoken content alone (Goldin-Meadow, 2009; So et al., 2009). Based on this hypothesis, speakers may use gestures for their own cognitive benefit (Kita, 2000; Krauss et al., 2000). While some experimental studies provide evidence supporting the trade-off hypothesis (Bangerter, 2004; Melinger & Levelt, 2004), others tend to favour the hand-in-hand one (De Ruiter et al., 2012; Krauss et al., 2000).

Recent research has delved into these two hypotheses with diverse objectives, and within both linguistic and non-linguistic domains (Cruz et al., 2017). In the linguistic context, previous studies have addressed these two hypotheses by looking at hand gestures and their frequency, but more recent investigations have also evaluated the hypotheses in the context of audiovisual prosody research (Cruz et al., 2017; Zygis & Fuchs, 2023). In a broader sense, the focal point of inquiry has been an examination of the interplay between verbal and visual prosody that aims to determine whether these modalities operate in parallel or complement each other. While some earlier studies discussed in the previous section (Borràs-Comes et al., 2014; Crespo-Sendra et al., 2013; Prieto et al., 2015; Rossano, 2010) found a trading relation between linguistic means of question-making across different languages and the extent to which facial movements are relied on to differentiate questions, others (Cruz et al., 2017) have emphasised how the auditory cues of prosody play a complementary role to visual cues.

Whether facial expressions enter a trade-off relationship with acoustic correlates of prosody or go hand-in-hand with speech may depend on the type of gestures, communicative setting and relevant constraints. Therefore, experimentally based evidence is required for more conclusive interpretations. Despite the earlier literature having frequently examined the relationship between gesture and voiced speech, there are still open questions that have not received a complete answer: what happens to gestures when fundamental frequency (F0), the most prominent prosodic cue, is absent from the acoustic signal, as is the case in whispered speech? Do gestures enter a trading relationship more intensively to enhance the perception of whispered speech? Due to the difference in the production mechanism of whispering, mainly the lack of phonation, the acoustic speech signals become voiceless and therefore the speech is harder to comprehend, which can be considered as a communicative constraint and may affect the degree of gesture production. Bar a scant number of studies, questions of this type have not been extensively addressed in the literature. Dohen and Loevenbruck (2008) noticed the compensatory effect of orofacial gestures in enhancing the perception of French whispered speech, in which acoustic signals were

degraded. The perception of focus in particular was significantly enhanced when audio and visual modalities were integrated in comparison with visual-only or audio-only conditions. Tao and Busso (2014) also investigated the role of the lips in recognising whispered as opposed to normal (voiced) speech, reporting that the combination of audio and visual features (lips) increased the accuracy of recognising whispered speech. Furthermore, Żygis et al. (2017) revealed that the articulation of vowels in whispered questions involved higher eyebrow raising and lip aperture than in statements compensating for the lack of F0, which supported the trade-off hypothesis. In a recent study examining the relationship between the acoustic signal and orofacial expressions in whispered versus normal speech and under visible versus invisible conditions, Żygis and Fuchs (2023) reported that the relationship between acoustic properties and gestures does not provide straightforward support for either a 'trade-off' or a 'hand-in-hand' hypothesis. When producing whispered speech, speakers may use more pronounced gestures and longer word duration to compensate for the lack of the fundamental frequency (supporting the trade-off hypothesis). On the other hand, since the gestures were also enhanced when the listener was invisible, the authors concluded that the orofacial movements were not produced solely for the needs of the listener (supporting the hand-in-hand hypothesis), but to help the speaker achieve an overarching communicative goal. Thus, investigations of whispered speech and (in)visibility condition may offer a unique opportunity to test the possible effects and verify the 'trade-off' versus 'hand-in-hand' hypotheses with respect to the relationship between speech and facial gestures.

In a similar vein, we will explore here whether facial expressions compensate for the absence of F0 and reduced amplitude, or whether they are redundant, by focussing on two communicative situations: one in which the speakers whisper compared with a situation where they use a voiced mode of speech, and another where they wear a protective face mask in comparison with a situation where they do not. In our analysis, we focus on eyebrow movements, lip aperture and eye opening as the target facial gestures. From a linguistic point of view, the variations in the intonation of questions (with rising pitch) versus statements (falling F0) will enable us to examine if facial gestures contribute to expressing these intonational differences between questions and statements. Moreover, we will be able to test whether facial gestures, in their probable interaction with the acoustic prosodic cues of sentence types, compensate for the degraded signals of whispered questions. Before proceeding to our experiment, we will first discuss the acoustic correlates of prosody in whispered speech and the effect of masks on speech communication.

1.4. Whispering and its prosody from an acoustic point of view

Whispering is one communication mechanism in human dialogue, used in various situations to convey linguistic information to listeners for different purposes (Tartter, 1989). As a paralinguistic phenomenon, it can be used in public and private domains with various purposes. From a social perspective, it may be adopted to restrict communication for reasons of privacy or secrecy (Frühholz et al., 2016), in which case the intelligibility of speech is to some extent lost due to the distortion of the speech signal (Jovičić & Šarić, 2008), or else in a situation where the speaker may wish to communicate clearly but not distract other likely audiences, for example, in a library. Apart from verbal information, whispering can also encode emotional cues

and evoke feelings of withdrawal or social isolation (Żygis & Fuchs, 2019). In pathological settings, it is observed in association with specific speech pathologies.

Despite the lack of the fundamental frequency (F0) in whispered speech, prosody is, though heavily dependent on F0, still discernible to some limited degree in whispered speech (Heeren & Lorenzi, 2014; Heeren & van Heuven, 2009). For example, studies have shown that listeners are able to distinguish sentence types in whispered speech expressed by various boundary tones where acoustic cues convey the key information (Fonagy, 1969; Heeren & van Heuven, 2009). It has also been reported that lexical tones can be perceived in whispered speech (Kong & Zeng, 2006; Liu & Samuel, 2004; Miller, 1961) and that listeners can differentiate intended pitch height as well (Higashikawa et al., 1996). The question that arises here is how information about intonational patterns for questions and statements is encoded in whispered speech.

While in the normal speech of several languages, F0 is strongly involved in the intonation of statements with a falling contour and polar questions with a rising contour, the task of executing intonation in whispered speech is delegated to segments. In previous research, many studies on whispered speech have sought vowel content, a number of them focussing on the acoustics of whispered consonants (Fan et al., 2011; Heeren, 2015a; Jovičić & Šarić, 2008); however, work on the prosody of whispering and its interplay with consonants has been relatively limited so far. In one of the few research papers on prosodic aspects of whispered speech, Żygis et al. (2017) investigated the correlates of utterance-final rising intonation in polar questions and falling intonation in statements in Polish. Their results showed that there are significant differences not only in the spectral properties of vowels but also in consonants when statements and polar questions are produced. That is, the intonation of questions was not only carried by a higher F1 and F2 in vowels, but also an increased intensity, greater spectral peak frequency, higher centre of gravity (CoG), higher standard deviation and lower kurtosis and skewness of consonants. Additionally, certain spectral dissimilarities between the production of questions and statements, encompassing spectral moments or slopes, were observed to a greater extent in whispered speech. This underscores the particular significance of these parameters in this speech mode. The authors proposed that the greater spectral disparities between questions and statements in whispered speech serve to compensate for the distinctive role of fundamental frequency (F0) in the normal speech mode.

The works examining vowels (e.g., Heeren, 2015b; Higashikawa et al., 1996; Higashikawa & Minifie, 1999; Meyer-Eppler, 1957) have acoustically attributed the perception of intonation to the formants, arguing that the first and second formants (F1 and F2) of whispered vowels are capable of carrying prosodic information. Some other studies have introduced intensity and duration as contributors of encoding pitch in whispered speech (Liu & Samuel, 2004; Meyer-Eppler, 1957). Kallail and Emanuel (1984) conducted an analysis of sustained vowels in male American English speech, revealing disparities in formant frequencies between whispered and normal speech. Specifically, their study found elevated values in F1 for whispered speech. Likewise, Higashikawa et al. (1996) observed higher formant frequencies in whispered /a/ compared with normally spoken /a/ among both male and female speakers of Japanese. Meyer-Eppler (1957) reported two acoustic correlates of the pitch in an examination of the German vowels /a, e, i, o, u/ on five tones on a diatonic scale and under two conditions: (a) singing the vowels in whispered mode and (b) whispering the statements or questions with a level falling or rising tone on the final syllable.

The results showed that the vowels /a, u/ had formant changes, particularly on the third formant (F3), whereas the vowels /e, i, o/ had increased noise when intended as higher. Higashikawa and Minifie (1999), in a study manipulating formant shifts in the first (F1) and/or second formant (F2) on synthesised whispered vowels, were able to demonstrate that the perception of pairs of vowels by listeners was more accurate when both formants were changed and when larger changes were applied. In a similar vein, Heeren (2015a) probed voiced and whispered Dutch vowels /a, i, u/ at three pitch levels – slow, middle and high –, noticing not only higher F1 and F2 but also a larger CoG in whispered vowels than in vowels produced in normal speech. Also, the differences between high versus low pitch levels and between high versus medium pitch targets were larger in whispered than in normal speech. The intensity was lower in whispered than in normal speech, and lower in the vowels /i/ and /u/ than in /a/. However, no systematic variation in the vowels' relative durations as a function of pitch target and speech mode was detected.

From a perceptual standpoint, researchers have also examined acoustic cues associated with whispered speech. An example can be found in the study by Higashikawa et al. (1996), which demonstrated that the differentiation between high and low /a/ in whispered speech can be perceptually discerned through differences in formant frequency. A study conducted by Heeren and van Heuven (2009) examined the identification of phrasal prosody, specifically in the context of whispered speech. The results indicated that the identification of questions and statements in whispered speech was slightly better than chance. Furthermore, the spectral tilt of vowels in the sentence-final syllable played a significant role in perceiving the sentence mode in whispered speech. Heeren (2015a) looked at listener sensitivity to consonantal cues to pitch in whispered versus normal speech. The findings indicated that in VCV (vowel-consonant-vowel) stimuli, discrimination accuracy was lower and processing speed slower in whispered than in normal speech. This disparity was attributed to variations in the available acoustic cues for listeners. However, when specifically examining fricatives in isolation, the processing speed and accuracy of pitch information were similar between the two speech modes, indicating that fricative cues play a subordinate role.

1.5. *The effect of face masks on speech*

In addition to obscured visual cues, acoustic attenuation can also influence a listener's ability to comprehend speech when a speaker is wearing a mask (Knowles & Badh, 2022; Pörschmann et al., 2020). The utilisation of face masks functions as a low-pass filter on speech, primarily due to their role as a physical barrier to the transmission of the acoustic signal. Face masks have been found to attenuate acoustic energy above the range of approximately 1–2 kHz (Corey, Mascola, et al., 2020; Palmiero et al., 2016). Corey, Jones, and Singer (2020) reports that face masks attenuate high-frequency sounds in front of the talker, with the strongest attenuation above 4 kHz. The study by Maryn et al. (2021) reveals that wearing masks significantly influences the acoustic markers relevant to clinical speech, including variations in fundamental frequency. Although masks have been observed to attenuate higher frequency components of the speech signal, there are variations in the observed impact of face masks on the reduction of speech intensity. For instance, a study by Fiorella et al. (2021) found that wearing a surgical mask did not lead to a substantial

reduction in speech intensity during the production of a sustained vowel. However, at the individual level, a majority of the participants demonstrated a decrease in speech intensity when wearing the mask, whereas a smaller portion exhibited an increase. The authors proposed that certain speakers may unconsciously exert greater vocal effort to compensate for the filtering effects of the masks. However, the acoustic analysis of infant directed speech by Cruz et al. (2022) showed that the mask affects the mean intensity of the speech signal. In another study, Cohn et al. (2021) found that sentences produced with a fabric mask exhibited higher descriptive mean speech intensities ranging from 0.1 to 2 dB SPL (decibel of sound pressure level) compared to those produced without a mask. This observation was consistent across three distinct speech styles of habitual, clear and emotional. The authors interpreted these findings as evidence that masks do not display a consistent pattern of intensity that distinguishes them from the absence of masks. Power distribution, spectral tilt and timing are other aspects of the acoustic signal affected by face masks (Rahne et al., 2021). Knowles and Badh (2022) observed alterations in spectral density characteristics, CoG and spectral variability (in habitual speech) and spectral tilt (across habitual, loud and clear speaking styles). KN95 masks demonstrated a greater effect on speech acoustics than surgical masks. The overall pattern of changes in speech acoustics was consistent across all three speech styles.

Based on the above-mentioned studies, face masks potentially obstruct not only the visual cues but also acoustic characteristics of speech, all of which are known to contribute to the perception of prosody in speech. In the present study, we extend recent research into the speech and the use of masks by examining the production of speech prosody in different speech modes and a face mask. In particular, we will examine intonational statement/question prosody produced with and without masks in whispered and voiced speech modes. Given that speakers' possibilities for expressing intonation in whisper are restricted on one side and the obstructive impact of the face mask on the other side, we examine if the probable reduction in acoustic and facial information affects speakers' production of sentence type prosody. We assume that acoustic cues may be altered, and facial gestures may be pronounced by speakers whispering behind their face mask to express intonation in comparison to the condition when they produce voiced speech without a face mask. Sinagra and Wiener (2022) found that masks make it more difficult to understand a speaker's intended intonation (question vs. statement) and that speakers may adapt their speaking, making their intonations explicit to the listener. In line with this, we test the possible compensation effect and the trade-off and hand-in-hand hypotheses in relation to the use of a face mask and orofacial gestures. We expect a mutual compensation effect across speech modes when a face mask is used, i.e., speakers with a face mask may use their facial expressions more intensively to enhance the perception of their intonation, especially when they whisper.

1.6. Individual differences in the use of orofacial gestures

Individual variations can be dependent on various factors, including cultural, psychological, cognitive, biological and neurobiological factors. There can also be gender-specific differences in the use orofacial gestures. For instance, in earlier research on nonverbal communication, women are generally acknowledged as more

proficient senders of nonverbal information compared to men (e.g., Buck et al., 1972, 1974; Dimberg & Lundquist, 1990; McDuff et al., 2017; Wallbott, 1988). Literature reviews consistently suggest that females are not only more facially expressive than males, but also tend to be more accurate in the recognition of facial gestures in general (Briton & Hall, 1995; Dimberg & Lundquist, 1990; Forni-Santos & Osório, 2015; Krumhuber et al., 2007; Wallbott, 1988). However, studies assessing the attributes involved in individual differences have been limited to the expression or recognition of emotional states. The focus of the present analysis is not to probe the individual variations and their underlying contributors. However, since this experiment is performed in a language in which the application of specific sociocultural norms may elicit differences in facial expressions between individuals, particularly between male and female speakers, we will specifically test the possible differences in orofacial expressions between these two groups of speakers. We believe that the results of the analysis may provide further insights into male versus female distinctions in orofacial gestures in non-emotional contexts and within the field of multimodal speech prosody research.

1.7. Predictions

The present study is an endeavour to enlarge the spectrum of research on the audiovisual properties of speech by investigating whether orofacial gestures contribute to the expression of intonational contrast between yes/no questions with rising intonation and statements with falling intonation under two communicative constraints – voiced versus whispered speech mode (which lacks F₀), and the use of face masks (which affects the acoustic signals of speech). Taking into account previous literature, we hypothesise that speakers will intensify their orofacial expressions when confronted with ‘marked’ conditions, that is, when they whisper, wear face masks and produce questions. In addition, speakers will employ the most pronounced facial expressions (eyebrow movements, lip aperture and eye opening) when they produce questions in whispered speech mode while wearing a face mask to compensate for the absence of F₀, reduced intensity, and lack of visibility of the lower face. We also predict that the employment of orofacial gestures will differ in two different styles – reading versus sentence imitation task – by being more pronounced in the sentence imitation task.

We will also assess acoustic cues signalling intonation of sentence type, including mean amplitude and word duration. We predict a longer duration of sentence-final words (i.e., the subject of our investigation) and reduced intensity in whispered speech mode. When interlocutors whisper and produce questions, the duration should be the longest (interaction). We predict a similar effect when speakers wear a face mask and produce questions. The reading mode should also exert longer duration and higher intensity due to possible hyper-articulation.

Furthermore, we will examine the extent to which the gestural and acoustic parameters correlate. The analysis of gestures in parallel with acoustic signals will allow us to test if facial gestures compensate for the absence of F₀, in which case the trade-off hypothesis will be confirmed. But if the intensity of orofacial gestures and acoustic parameters prove to be positively correlated, that would confirm the hand-in-hand hypothesis, suggesting that in the ‘marked’ conditions these gestures are primarily produced to fulfil the speaker’s own communicative needs. Due to the

absence of F0 in whispered speech, our analysis will focus solely on examining the mean intensity and duration (see below for details).

In addition, we will test the orofacial movements in different speech styles excerpted through a reading and a sentence imitation task (see Section 2.2 for details). We hypothesise that the imitation task performed in a dialogue with a confederate will elicit more intensified facial expressions as opposed to the reading task.

To summarise our hypotheses:

- (1) Speakers use more pronounced facial expressions, that is, higher eyebrows, with larger lip and eye openings
 - a) in whispered speech in comparison to normal speech,
 - b) with as opposed to without a mask,
 - c) in questions rather than statements,
 - d) in questions produced in whispered speech while wearing a face mask (a three-way interaction).
- (2) The duration of a sentence-final word will be longer
 - a) in whispered speech in comparison to normal speech,
 - b) with as opposed to without a mask,
 - c) in questions rather than statements,
 - d) in questions produced in whispered speech while wearing a face mask (a three-way interaction).
- (3) The normalised mean amplitude will be larger
 - a) in normal than in whispered speech,
 - b) with as opposed to without a mask,
 - c) in questions rather than statements,
 - d) in questions produced in normal speech without a face mask (a three-way interaction).
- (4) Based on the trading hypotheses, we expect
 - a) a negative correlation between acoustic parameters (intensity, duration) and orofacial expressions if the trade-off hypothesis holds true;
 - b) a positive correlation between acoustic and orofacial expressions if the hand-in-hand hypothesis holds true.
- (5) Male and female speakers will display difference in the use of gestural parameters, that is, female speakers will produce more intensified pro-facial expressions than male speakers
- (6) Speakers will show more intensified orofacial expressions in the imitation task than in the reading task.

2. Methodology

2.1. Participants

The facial movements in this study were collected during video recording sessions. Ten native speakers of Persian, five males and five females (mean: 30.6, SD: 5.03, age range 20–35), were recruited to participate in the experiment. They self-reported normal vision and hearing, with no history of speech impairments. All participants were financially compensated for their time. They filled in a short demographic questionnaire and provided written informed consent to the study protocol.

2.2. Experimental design and stimuli

Speakers' facial movements were recorded in two phases:

1. *Reading* – where the participants were instructed to read a series of questions and statements. Each sentence was displayed on a monitor positioned in front of them.
2. *Sentence imitation task* – where the participants interacted with an interlocutor to produce questions and statements. In this phase, the confederate – the same speaker throughout the whole experiment – generated either a question or statement in a voiced or whispered speech mode, and the participants were supposed to respond to the question by converting it into a statement or ask a question in response to the statement by altering their intonations in the same speech mode. See examples below:

(1) **Question**

Confederate: Diruz Kusha goft sepid.

'Yesterday, Kusha said white'.

Informant: Diruz Kusha goft sepid?

'Yesterday, Kusha said white?'

(2) **Statement**

Confederate: Diruz Sasan goft kabab?

'Yesterday, Sasan said kebab?'






Informant: Diruz Sasan goft kabab.

'Yesterday, Sasan said kebab'.

Each phase of the experiment was conducted in two different settings related to speech mode (normal vs. whispered speech) and mask (wearing a KF94-3D face mask vs. without a mask, see Table 1). Thus, four stimuli blocks were designed, each containing a list of sentences (questions vs. statements). In each block, three conditions were under the focus of this study: (a) *speech mode (modal, whisper)*, (b) *mask condition (with mask, without mask)* and (c) *sentence type (statement, question)*. In both phases of the reading and sentence imitation tasks, the presentation order of blocks as well as the order of sentences within these blocks were randomised for each speaker to prevent order effects.

The stimuli consisted of 20 Persian sentences or 10 pairs of statements and questions (see [supplementary material](#) for the full list of sentences). The selected sentences were all the same, differing only in the meaning of the target words at the sentence-final position and punctuation, that is, statements ended with a full stop and questions with a question mark. It should be noted that we did not add distractor

Table 1. Blocks of stimuli

		Speech type	
 Without a face mask	Whisper		Modal 
	Whisper		Modal 

sentences on purpose because our aim was to exert intonational differences between questions and statements, which, in the case of whispered speech and mask condition, is a challenging task per se. Adding distractor with other intonation contours/sentence types would, in our view, make the relatively long experiment unnecessarily longer and more demanding for participants.

The target words for the experiment were selected based on the following criteria: all of them were bisyllabic content words with the stress falling on the second syllable. The second syllable of all the words had a CVC structure, starting with a bilabial stop /p/, /b/ or /m/ and followed by unrounded vowels /a/, /ɛ/ or /i/ (for instance, 'sepah' [se'pəh] meaning 'army', 'tapes̄h' [tæ'peʃ] meaning 'beat' or 'sep̄id' [se'pid] meaning 'white'). The purpose behind the inclusion of bilabial stops was that they involve lip closure in their articulatory realisation, prior to a lip aperture for the following vowel, thus facilitating the subsequent measurements. Also, the presence of a stop phase in plosives facilitated acoustic annotation. The vowels are assumed to vary in their height, from the greatest lip aperture for /a/ to the smallest one in the case of /i/.

2.3. Experimental setting and recording procedure

The recordings were obtained in a soundproof studio in Tabriz, the provincial capital of East Azerbaijan, northwestern Iran. Participants were seated and asked to position their heads at the centre of a frame with a solid green uniform background, situated about one metre from a tripod-mounted video camcorder (Sony Alpha a6400 Mirrorless Camera with 16–50 mm lens). A lightweight field monitor (VILTROX DC-70 II 4K HDMI, 7-inch TFT high-resolution LCD panel), connected to a portable computer, was clipped onto the camcorder and displayed the stimuli so that in the *reading phase* of the experiment, the participants could simultaneously look at the lens of the camera and read the stimuli on the screen. In addition to the video recordings, auditory data were synchronously captured using a Zoom H6 APH recorder with a 120-degree microphone connected to the camcorder through a standard stereo cable sampled at 44.1 kHz, digitised mono. The recorder was mounted on a tripod and appropriately located 15 cm from the participants. In order to prevent head movements and ensure a fixed distance to the lens of the camera, the participants were instructed to try to avoid body or head movements. [Figure 1](#) illustrates the experimental setup to record the two phases of the reading and sentence imitation tasks. The participants' written consent was obtained to use their faces for scientific publication purposes.

Prior to the experiment, the procedure of the experiment was explained to the participants. In the *reading phase*, they were supposed to produce each sentence in time with its display on the monitor. In the *sentence imitation task*, the participants were given time to practise a few samples (two to four sentences), converting the questions into statements and vice versa, ensuring they were able to produce a prosodic contrast between questions and statements. Once the participants felt ready, they were recorded. As already explained, the experiment was conducted in two phases. In the first phase, each participant started by reading four lists of randomised sentences (questions and statements) either in whispered or voiced speech and with or without a face mask (see [Table 1](#) for details). After completing each list, the participants would take a pause and were then asked to start the next unit of sentences. After a break, the second phase of the experiment, sentence imitation

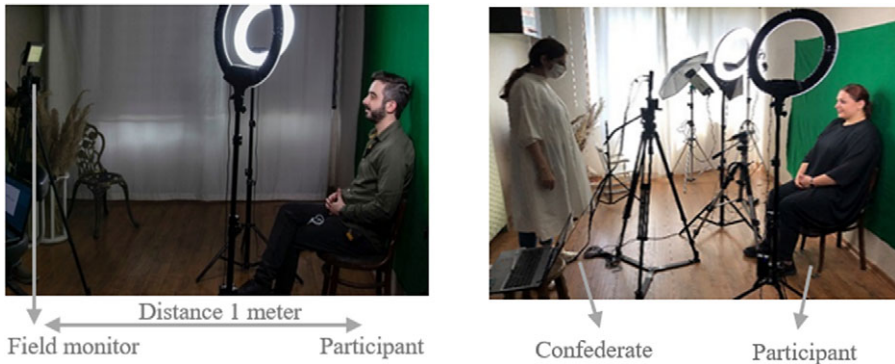


Figure 1. Experimental setup. Reading phase of the experiment: field monitor displays the sentences one by one (left). Sentence imitation phase of the experiment (right): confederate and the participant exchanging the questions and statements in voice mode of speech (the data were collected in the peak of COVID-19, so to avoid the risk of being infected by the virus; the confederate wore a surgical mask during the whole experiment).

task, was performed between the participant and a confederate. During the experiment, the researcher made sure that questions and statements were produced in a proper manner by each participant. In both phases, the sentences across conditions were identical; however, the order of presentation differed in each block. In total, 80 sentences were recorded for each speaker in each phase.

2.4. Acoustic annotation

Using *Praat 6.0.40* (Boersma & Weenink, 2018), we semi-automatically annotated the following units (see [Figure 2](#), where the annotation of a sample sentence in voiced speech mode is exemplified):

- a) *sentence*: the beginning and offset of the sentence (A in [Figure 2](#));
- b) *final word*: the onset of the sentence-final word and its offset, corresponding to the end of the sentence (B in [Figure 2](#));
- c) *unstressed syllable*: the onset of the first syllable of the sentence-final word and its offset (C in [Figure 2](#));
- d) *stressed syllable*: the onset of the second syllable of the sentence-final word and its offset (D in [Figure 2](#));
- e) *vowel*: the onset and offset of the vowel in the stressed syllable (E in [Figure 2](#)).

The following acoustic parameters were measured:

- *mean amplitude difference* (db) (mean amplitude of the word final syllable – mean amplitude of the word-prefinal syllable);
- *sentence-final word duration*;
- *maximum F0 for each of eight equal intervals of a sentence*.

We calculated the *mean amplitude difference* in order to better control the mouth-to-microphone distance and gain more information about the normalised amplitude of

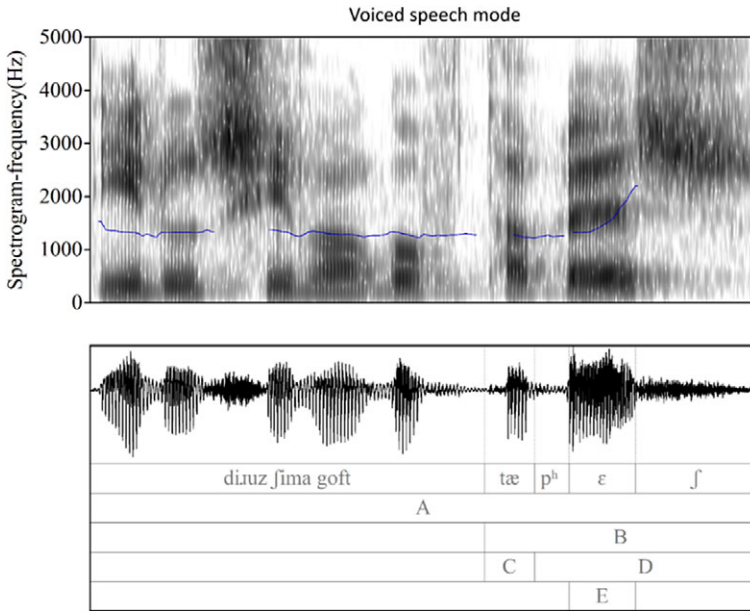


Figure 2. Spectrogram and oscillogram for voiced speech with annotation labels for the sentence ‘Diruz shima goft tapesesh?’ (‘Yesterday, Shima said “beat?”’). Blue line marks F0 in voiced speech. Alphabet letters represent the annotated intervals.

the target unit. We also calculated the maximum F0 in voiced speech with and without a face mask over all the sentences to ensure that the maximum F0 peak was higher in the sentence-final items in questions than in statements. Prior to the analysis, the sentence duration was divided into eight equal intervals, and the maximum F0 was excerpted for each interval.

2.5. Analyses of orofacial movements

To pursue our research goal, we tracked and quantified the movements of different face structures by utilising a facial landmark detector. We designed a pipeline which used a feature combination of two python libraries to achieve accurate facial landmark detection in the video recordings: OpenCV (2015), a library of python binding; and a facial landmark detector inside the *Dlib* library (King, 2009) containing a pretrained machine learning model based on histogram of oriented gradients and linear support vector machine method. By means of these two open-source tools, videos were iterated frame by frame at 29.969 frames per second. In every frame, the face region was identified and 68 2D landmarks mapped to distinctive units of the face. Then, the estimates of x, y coordinates for each landmark were extracted in pixels. Figure 3 represents the indexes of the 68 coordinates on a participant’s face. Also, the orofacial expressions in a question versus a statement with and without a face mask are illustrated in Figure 4 (participant’s written consent was obtained to use his photos for publication).

It should be noted that the landmark detector did not accurately locate the landmarks on the eyebrows and eyes in some video files recorded of two participants

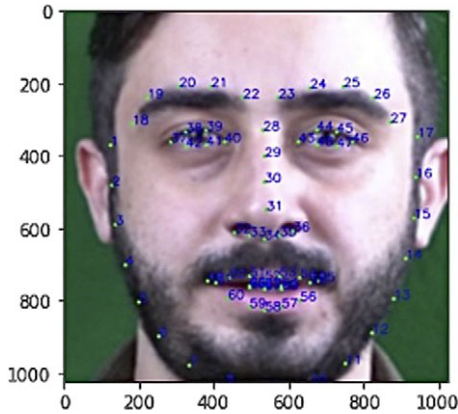


Figure 3. Sixty-eight facial landmarks tracked by OpenCV.



Figure 4. Expression of a statement (left) and a question (right) by a speaker with and without a mask.

wearing face masks. We therefore excluded these files in our analysis. Each video file contained one produced sentence, so from 1600 sentences overall, 143 were excluded. From each video file we excerpted the sentence-final word for further analysis.

As an important requirement, we normalised the input video frames to minimise the likely unexpected head movements and scale orientation of the face in each frame.

Table 2. Face vectors used for the analysis

Face vectors	Calculations based on the facial landmarks
Eyebrow raising	D1 (distance between mean of points 19, 20, 21 and 28 for the right eyebrow and mean of points 24, 25, 26 and 28 for the left eyebrow)
Eye opening	D2 (distance between points 38 and 41 for the right eye and points 45 and 48 for the left eye)
Lip spreading	D3 (distance between points 49 and 55)
Lip aperture	D4 (distance between mean of points 51, 52, 53 and mean of points 59, 58, 57)

To this end, we used an affine transformation by which the distance ratio between the centre of each eye, as well as the angle between the eye centroids, was calculated. The midpoint between the left and right eyes, atop the nose, was taken as a reference position serving as the (x, y) coordinate, in which the face was rotated so that the eyes lay along the same y coordinates. With this method, a canonical representation of the face was obtained.

We defined four face actions based on the key points of the face, including eyebrow raising, eye opening, lip spreading and lip aperture (see Table 2 and Figure 5 for an overview). Once the x, y positions of facial landmarks were derived, the Euclidean distance (*D*) between the key points corresponding to each face action was calculated for each successive frame as presented in (1):

$$D = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \tag{1}$$

The average summation of distances (*D_n*) was calculated for each face action vector to obtain the absolute mean value of movement within the given recording window (pixel/frame), as shown in (2):

$$\text{Mean movement of face vector} = D_{(n1)} + D_{(n2)} + D_{(n3)} + \dots / fnum \tag{2}$$

where *n* is the number of each frame per video and *fnum* is the total number of frames per video.

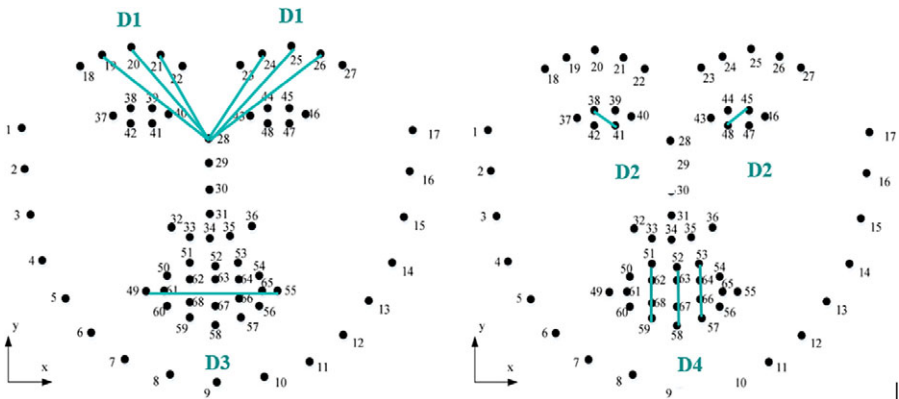


Figure 5. Sixty-eight facial landmarks tracked by OpenCV and Dlib, with four face vector distances (image from the iBUG 300-W dataset by Sagonas et al., 2013).

Finally, the resulting speed vectors were low-pass filtered by means of a Savitzky–Golay filter, then parameterised with a window size of seven samples and a third-degree polynomial. This smoothing helped us to iron some of the jitter inherent to video-based tracking, without affecting the main shape of the signal (Pouw & Trujillo, 2021).

2.6. Statistical analysis

We conducted the statistical analyses in R studio 4.2.2 (RStudio Team, 2022) using the packages *lme4* (Bates et al., 2020) and *emmeans* (Length, 2019). Applying linear mixed models, we assessed the effects of speech mode (*normal*, *whispered*); mask condition (*with mask*, *without mask*); sentence type (*question*, *statement*), vowel type (*a*, *ε*, *i*); gender (*male*, *female*) and speech type (*reading*, *sentence imitation task*) on the following dependent variables: left eyebrow raising, right eyebrow raising, eye opening, lip aperture and lip spreading, as well as the acoustic variables of word duration and mean amplitude difference. The reference levels in all these models were ‘*whispered speech*’, ‘*without mask*’, ‘*statement*’, vowel [*a*], ‘*male*’ and ‘*sentence imitation task*’. We also included a three-way interaction of speech mode, mask condition and sentence type to test whether the dependent variables are most extreme in the most complex condition, that is, when speakers whisper, wear a face mask and produce questions. A random structure was added to the models: word and participant as random intercepts, and by-participant and by-word slopes for the speech mode, mask condition, sentence type, speech type and vowel.

The complex design of models often led to non-convergence issues. In these cases, some random slopes were removed after examining their correlations (e.g., 1 or -1) (Winter, 2020). It should also be pointed out that normal distribution of residuals in all the models was checked before running the final models. Finally, we used the *emmeans* () function with Tukey adjustment to perform pairwise comparisons of mean values of the dependent variables under different conditions.

In addition, we examined the extent to which the gestural and acoustic parameters correlated normalising all the continuous variables of the data by z-scoring them for each speaker independently. Word duration and the mean amplitude difference are the two acoustic parameters selected as the predictors of each orofacial gesture. F0 was excluded as it did not appear in whispered speech. We also included a by-speaker and by-word random slopes for each predictor. An example is given in (3).

$$\begin{aligned} \text{lmer}(\text{z_scored_lip_aperture} \sim \text{z_scored_word_duration} \\ + (1 + \text{z_scored_word_duration} \mid \text{participant}) \\ + (1 + \text{z_scored_word_duration} \mid \text{word, data} = \text{data}). \end{aligned} \quad (3)$$

The initial and final models for each dependent variable and their outputs as well the results for pairwise comparisons of mean values are presented in the [supplementary material](#).

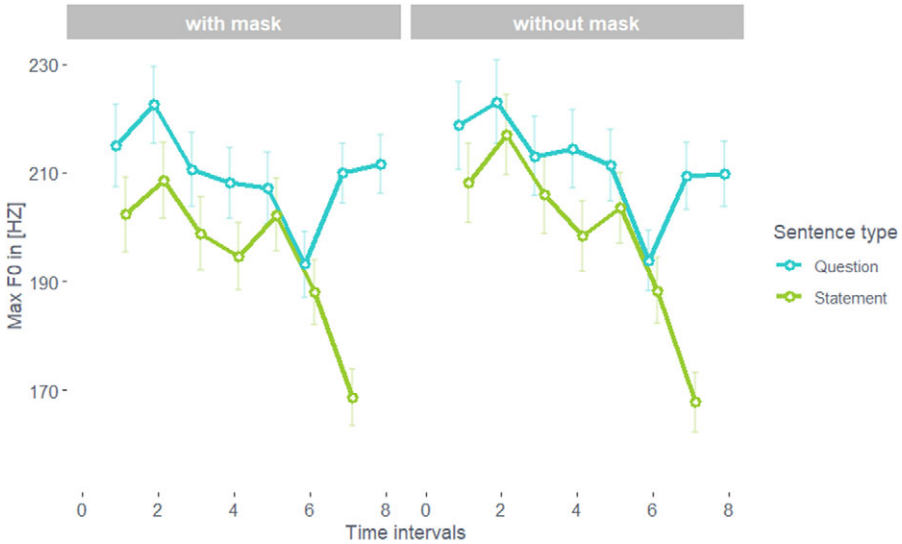


Figure 6. Maximum F0 values calculated for eight intervals across all questions and statements in voiced speech with and without a face mask. The dots show mean values, and the whiskers show 95% confidence intervals.

3. Results

Before presenting our results, we wanted to check whether the prosodic patterns of questions versus statements are different by measuring the F0 in voiced speech mode to confirm that the speakers had used the expected patterns in the experiment. To this end, we calculated the maximum F0 in voiced speech over all the sentences, which were divided into eight equal time intervals prior to the analysis. The results (see [Figure 6](#)) show a clear difference in F0 between questions and statements: F0 is higher in the second part of the question as compared to statement confirming that speakers successfully realised the respective intonation of question or statement.

3.1. Orofacial gestures

3.1.1. D1: Eyebrow raising

Based on the results of linear mixed-effect models, the left eyebrow is more raised in whispered than in normal speech ($t = 4.416$, $p < .001$), in questions than in statements ($t = 5.494$, $p < .001$), and when wearing a facial mask ($t = 4.705$, $p < .001$). The interaction *Speech Mode* \times *Sentence Type* \times *Mask Condition* did not turn out to be significant. However, when we compare normal versus whispered mode of speech in mask and without mask condition, it turns out that the interaction between *Speech Mode* and *Mask Condition* is significant with the highest eyebrow raising found in whispered speech with mask (*Speech Mode* \times *Mask Condition*, $t = 3.505$, $p < .001$, see [Figure 7](#) left). Finally, a significant interaction between the *Speech Mode* \times *Sentence Type* shows that questions in the whispered mode of speech are produced with a higher raising of the left eyebrow than in normal speech. This difference for statements is smaller ($t = 3.344$, $p < .001$, see [Figure 7](#) right). Note that most of the pairwise comparisons were significant for

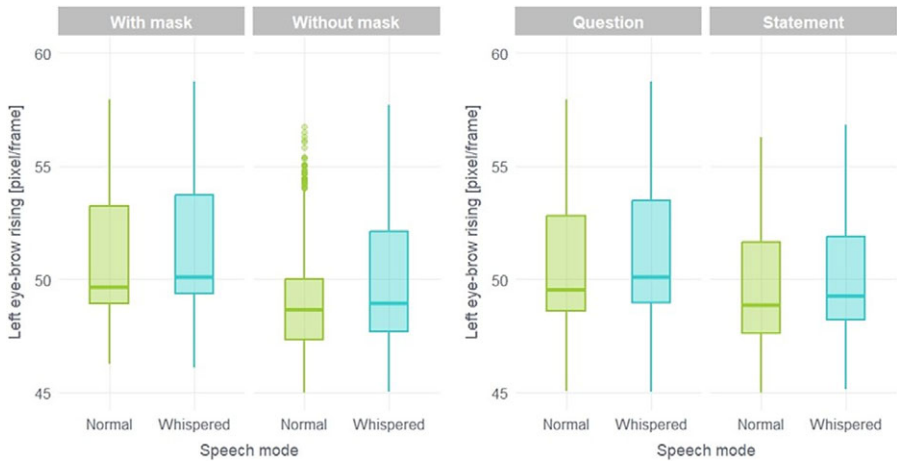


Figure 7. Left eyebrow raising in the sentence-final word: interaction between **speech mode** and **mask condition** (left) as well as **speech mode** and **sentence type** (right).

both interaction types. See the [supplementary material](#) for more details. As far as the effect of gender is considered, we did not find a significant difference between male and female participants raising their left eyebrows. Similarly, there was no significant difference in left eyebrow movement between reading and sentence imitation speech styles.

Regarding the right eyebrow, our results show it is more raised in the whispered mode than in normal speech mode ($t = 2.478, p < .05$), and in a condition when a face mask is used ($t = 3.957, p < .01$). Questions are also produced with a higher raising of the right eyebrow when compared with statements ($t = 5.235, p < .001$). Although the three-way interaction between sentence type, speech mode and mask condition was not significant, the statistical modelling reveals a significant interaction between speech mode and sentence type (*Speech Mode* \times *Sentence Type*, $t = 2.208, p < .05$) indicating a larger difference of eyebrow raising between whispered versus voiced mode of speech when questions are produced, see [Figure 8](#) left. Also, the significant interaction between speech mode and mask condition (*Speech Mode* \times *Mask Condition*) shows that the difference in right eyebrow movement between whispered and normal speech is smaller when a face mask is not used as opposed to the condition where a face mask is used ($t = -6.375, p < .0001$) ([Figure 8](#) right). Most of the results of pairwise comparisons across various conditions were significant, please see [supplementary materials](#). Finally, a significant difference between male and female speakers was found, with female speakers raising their right eyebrow higher than male speakers ($t = 2.832, p < .05$). Concerning the probable difference between reading versus sentence imitation speech styles, we did not detect a significant result.

3.1.2. D2: Eye opening

As with eyebrow movements, our results reveal a significant influence of speech mode on eye opening: both eyes are opened larger in whispered than in normal speech ($t = 3.628, p < 0.01$ for the left eye, and $t = 2.827, p < .05$ for the right eye). They

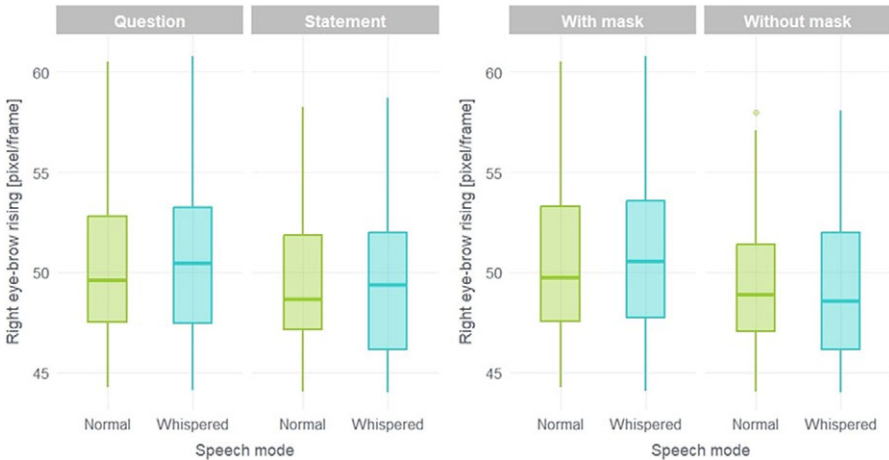


Figure 8. Right eyebrow raising in the sentence-final word: interaction between **speech mode** and **sentence type** (left) as well as **speech mode** and **mask condition** (right).

are also larger when producing questions as opposed to statements ($t = 6.864$, $p < .0001$ for the left eye, and $t = 5.813$, $p < .0001$ for the right eye). However, the results exhibited no significant effect for any of the interactions between the target conditions, that is, speech mode, mask condition and sentence type. Based on the results of statistical modelling, it seems that female speakers open their right eye wider than do male speakers ($t = 2.514$, $p < .05$), though the difference was not significant for the left eye. The comparison of reading versus sentence imitation task showed no significant difference between two speech styles.

3.1.3. D3: Lip spreading

To measure the movements of lips, the mask condition was excluded. Neither of the conditions (*Speech Mode and Sentence Type*) exhibited a significant effect for lip spreading. However, our results show an expected effect, that is, a wider spread of lips for the vowel [i] than for [a] ($t = 4.939$, $p < .01$) and [ɛ] in comparison to [a] ($t = 5.127$, $p < .01$). We found a significant difference between two speech styles, with a wider spread of lips in reading than sentence imitation task ($t = -4.843$, $p < .001$). Also, a significant difference between males and females was that male speakers have a wider spread of their lips ($t = 2.564$, $p < .05$).

3.1.4. D4: Lip aperture

Lip opening was significantly larger during whispered speech compared to normal speech ($t = 5.788$, $p < .001$) and in questions compared to statements ($t = 5.423$, $p < .001$). The significant interaction *Speech Mode × Sentence Type* ($t = -3.667$, $p < .001$) is reflected in a larger difference between whispered and normal speech modes for questions than for statements (see Figure 9). The significant results of pairwise comparisons are provided in the [supplementary material](#). The results exhibited neither a significant difference between male and female participants, nor between two speech styles.

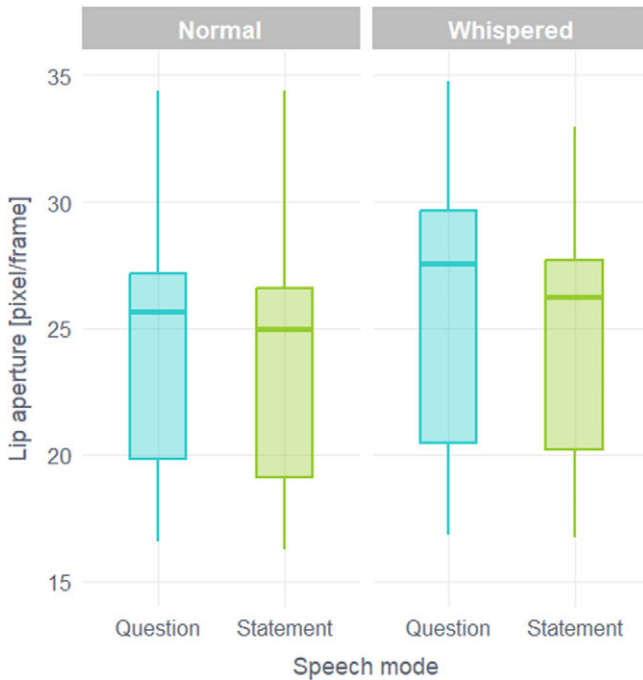


Figure 9. Lip aperture in the in sentence-final word: interaction between **speech mode** and **sentence type**.

3.2. Acoustics

As expected, sentence-final word durations were longer in whispered than in normal speech mode ($t = 3.688$, $p < .01$) and in questions than in statements ($t = 7.544$, $p < .0001$). We did not find a significant interaction between the three conditions of speech mode, mask condition and sentence type. But our results indicate that the difference in word duration between whispered and voiced speech is smaller when a face mask is not used. In other words, whenever speakers whisper with a protective face mask, they make the target words longer than they do without a face mask. This is reflected in a significant interaction of *Mask Condition* \times *Speech Mode* ($t = -3.585$, $p < .001$, see Figure 10 left). Similarly, the interaction between *Speech Mode* \times *Sentence Type* ($t = 2.229$, $p < .05$) shows a larger difference between whispered and normal speech when statements are produced, see Figure 10 right. However, for both interaction types, some of the pairwise comparisons were not significant, please see [supplementary material](#). Finally, the difference between male and female speakers is significant. Male speakers produced longer sentence-final words than did their female counterparts ($t = 2.947$, $p < .05$). The results did not exhibit a significant difference between reading versus sentence imitation task.

We also compared the mean amplitude (db) difference between the final (stressed) syllable and prefinal (unstressed) syllable of the target word. As expected, there was a marked difference between the mean amplitude in whispered versus normal speech, with a lower mean amplitude in the whispered mode of speech ($t = -4.657$, $p < .001$). It was also higher in questions than in statements ($t = 5.869$, $p < .0001$).

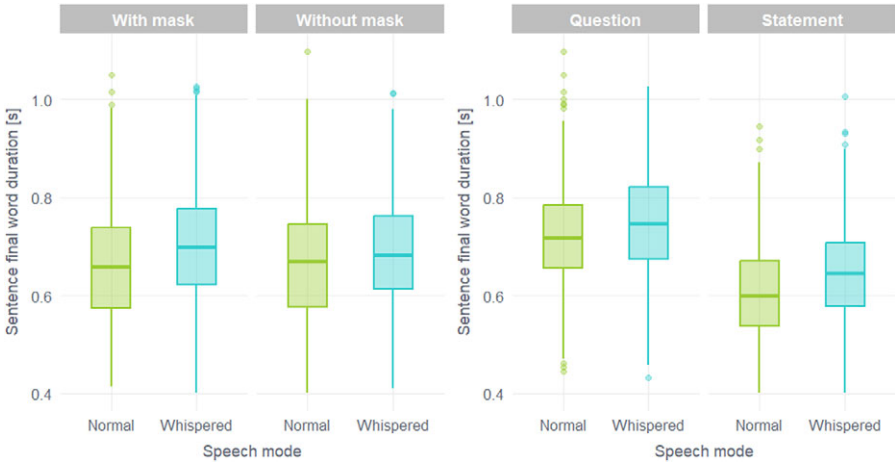


Figure 10. Duration of sentence-final words: interaction between **speech mode** and **mask condition** (left) as well as between **speech mode** and **sentence type** (right).



Figure 11. Mean amplitude difference between unstressed and stressed syllable of the sentence-final words.

The significant interaction between speech mode and sentence type (*Speech Mode* × *Sentence Type*, $t = -5.299$, $p < .0001$) reveals that the mean amplitude of questions produced in voiced speech as opposed to whispered speech is higher than the same difference found for statements, see Figure 11. Moreover, we found a higher mean

amplitude for vowel [a] than for [ɪ] ($t = 2.762, p < .05$). The difference in the mean amplitude between the vowels [a] and [ɛ] were not significant.

3.3. The correlation between orofacial and acoustic parameters

So far, the results obtained have analysed orofacial expressions and acoustic parameters separately. In the following section, we investigate the extent to which gestural and acoustic parameters exhibit a linear relationship. We selected *word duration* and *mean amplitude difference* as two acoustic parameters to examine whether there is a correlation between them and *left/right eyebrow raising*, *eye opening*, *lip aperture* and *lip spreading*. As stated earlier, all continuous variables were z-scored for each speaker.

The analysis of z-scored word duration shows that word duration does not correlate with the raising of the left eyebrow, but does with right eyebrow motion ($t = 3.663, p < .001$). A positive correlation was also detected between word duration and lip aperture ($t = 3.196, p < .01$, see Figure 12). It appears that when speakers produce a longer word, they realise a larger lip aperture. Similarly, we found a significant effect of word duration on left and right eye opening ($t = 3.348, p < .001$ for the left eye and $t = 2.836, p < .01$ for the right eye; see Figure 12).

In contrast to those for word duration, the results of statistical analyses show no significant correlations between mean amplitude difference and orofacial expressions.

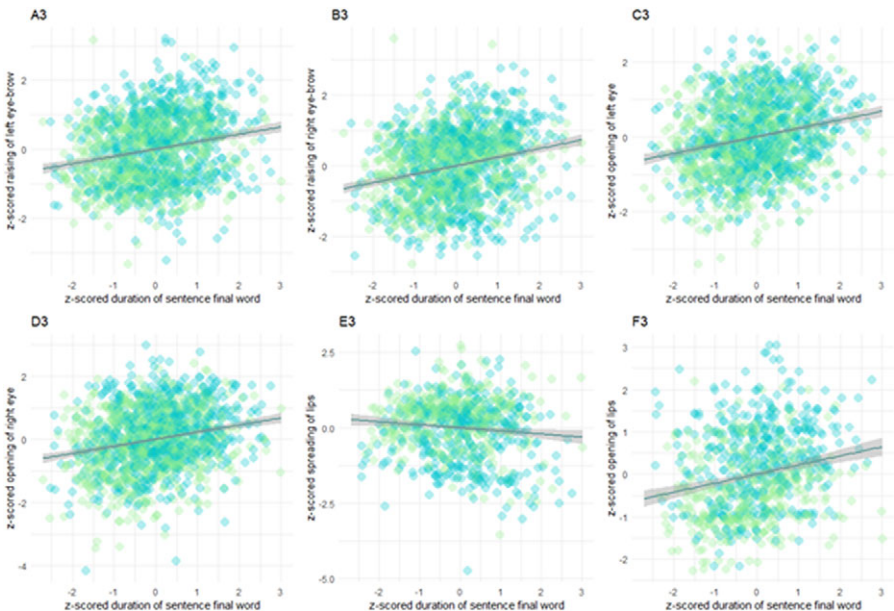


Figure 12. Scatterplots with z-scored **word duration** (x-axis) and **orofacial parameters** (y-axis). Linear regression lines ($y \sim x$) are given in dark blue (colour online) and confidence bands in grey. Data points corresponding to normal speech are visualised in light green, while data points for whispered speech appear in turquoise blue.

4. Discussion and conclusions

This study was an endeavour to investigate the multimodal prosody of speech under different communicative constraints. One of our main goals was to examine the potential interaction between orofacial expressions and acoustic signals under two communicative situations: (i) when speakers whisper or speak in normal speech mode and (ii) when speakers do or do not wear a face mask. To control for prosodic aspects, we introduced to our design polar questions, which are typically produced with rising intonation in Farsi; and statements, which are typically produced with a falling one. Our core question was whether there are any differences in the employment of orofacial gestures in statements and questions under restricted communicative conditions – that is, when fundamental frequency is absent from the acoustic signal and the intensity is reduced (as is the case with whispered speech), and when face masks are used, introducing changes to the acoustic signal and hiding a part of orofacial gestures. We also addressed the differences in the production of orofacial expressions between male and female speakers, and between the two speech styles of the reading versus sentence imitation task.

As hypothesised, the findings revealed more intensified orofacial expressions – that is higher raising of the eyebrows – as well as larger opening of the lips and eyes in the whispered mode of speech than in voiced speech, and in questions than in statements. Our findings provided some support for the hypothesis proposing an increase in the use of facial expressions when speakers wear a face mask as a greater eyebrow raising was observed when a face mask was worn, with the effect being larger in whispered than in normal speech. Furthermore, in line with previous studies showing an association between eyebrow movements, lip aperture and question marking (see, e.g., Cavé et al., 1996; Borràs-Comes et al., 2014; Torreira & Valtersson, 2015; Cruz et al., 2017; Žygis et al., 2017; Miranda et al., 2021; Sardhaei et al., 2022; Nota et al., 2021; Hömke et al., 2022; Žygis & Fuchs, 2023), we found higher raising of eyebrows and more intense lip aperture, with the effect larger in questions than in statements. Therefore, this study not only replicates earlier findings, but also extends prior research by demonstrating how various facial signals can play a supportive role in conveying the intonation of sentence type under specific communicative constraints such as whispered speech. Specifically, our findings provide evidence of language-specific differences in using orofacial expressions to mark questionhood. According to earlier literature (Borràs-Comes et al., 2014; Borràs-Comes & Prieto, 2011; Crespo-Sendra et al., 2013), there is a relationship between question marking strategies and the use of gestures to distinguish the sentence type meaning. Also, in some languages or language varieties, a relationship between pitch accent types and gesture types has been observed, which can be mediated by sentence type and/or pragmatic meaning (Cruz et al., 2017). Considering the fact that in Farsi, questionhood is marked by prosodic cues, our results are in line with previous research proposing the idea that in languages that lack morpho-syntactic question marking or informative auditory/tonal features to contrast the two distinct pragmatic meanings, visual modality can compensate for the absence by playing a trade-off role in expressing questionhood. However, when languages do have such marking strategies, visual cues can play a secondary role in distinguishing between statements and questions. Earlier studies by Srinivasan and Massaro (2003) on English and House (2002) on Swedish, demonstrated this by revealing a larger role for auditory than for visual cues in judging whether an utterance is a statement or question. On the other

hand, Crespo-Sendra et al. (2013) showed that Catalan listeners relied more on facial cues while Dutch listeners attended to prosodic ones, which was linked to the type of intonational cues employed, wherein Catalan speakers utilised the same intonational contour for both pragmatic meanings, distinguished by a varying pitch range. In contrast, Dutch speakers employed distinct intonational contours for the same purposes. Also, Borràs-Comes et al. (2014) found that Catalan speakers used more visual cues than Dutch speakers did. In an analysis of echo questions and statements in Brazilian Portuguese (Miranda et al., 2021), although listeners relied more on auditory cues than visual ones when presented under clear acoustic conditions, the visual channel had a higher beneficial effect when stimuli consisted of degraded auditory cues. This is consistent with Massaro and Cohen's assumptions (Massaro & Cohen, 1983) regarding the greater contribution of the visual channel when either the verbal information is ambiguous or the auditory information degraded (in our case whispered speech with degraded F0 and reduced intensity).

Furthermore, our results demonstrate major interindividual differences in the employment of orofacial expressions (see [supplementary material](#)). While such variation can arise from several factors such as cultural background, personality traits, gender and individual differences in facial anatomy and muscle control, we focussed on the gender of the participants. In particular, our results suggest that male speakers of Farsi tend to have a broader spread of lip movement, which could be attributed to differences in their facial anatomy. By contrast, our data show that females speakers of Farsi tend to use the right eyebrow and right eye more prominently than their male counterparts, which can be linked to the fact that females tend to produce more pronounced facial expressions, as suggested by earlier research (Dimberg & Lundquist, 1990; Hess & Fischer, 2013). This finding may appear unexpected since eyebrows are typically expected to move in synchrony. However, previous studies (Cavé et al., 1996; Žygis & Fuchs, 2023) have also documented variations in eyebrow movement, indicating that such differences are not unprecedented. Given the fact that different muscles drive eyebrow movements, we might expect different reactions for each eyebrow. The motor cortex, specialised in voluntary movements and with its various regions controlling specific muscles, likely influences eyebrow movements. Variations in the activation or connectivity of these specialised regions within the motor cortex could potentially result in discrepancies in left and right eyebrow movements. Likewise, Sato and Yoshikawa (2007) suggest that eyebrow movements may not occur in parallel due to the involvement of different brain hemispheres in perceiving and producing facial expressions. The lateralised organisation of the brain could underlie the asymmetry observed in left and right eyebrow movements (see, e.g., Cavé et al., 1996; Ekman et al., 1980; Hellige, 1993).

We selected word duration and mean amplitude difference as two acoustic parameters to see if they contribute to the production of sentence-type intonation under the target communicative constraints, that is, when speakers whisper wearing a face mask. Our results lend support to the hypothesis that speakers produce longer words when they whisper and produce questions, but the difference between the conditions of wearing and not wearing a face mask was not significant, which is in line with (Georgiou, 2022) reporting the minimal effect of face masks on the temporal aspect of speech. However, our research reveals that words were longer in the whispered speech mode when participants wore face masks. This lengthening effect

on word duration could largely be attributed to the cumulative impact of whispering while wearing a face mask.

The mean amplitude difference was higher in questions and lower in whispered speech – and highest in voiced mode questions. Our findings revealed no noticeable change in the mean intensity difference when participants wore face masks as opposed to the condition where they did not. In line with our results, existing studies on the effects of face masks on speech amplitude also present varying conclusions. While some suggest that face masks may cause a reduction in speech intensity (Cruz et al., 2022; Goldin et al., 2020; Rahne et al., 2021), the overall consensus is that they do not uniformly result in lower speech intensity (Cohn et al., 2021; Maryn et al., 2021; McKenna et al., 2022). For instance, some earlier research (Joshi et al., 2021; Lin et al., 2021; Magee et al., 2020; McKenna et al., 2021) comparing mask and no-mask scenarios shows a significant increase in vocal intensity when a mask is used. On the other hand, Nguyen et al. (2021) did not observe significant changes in vocal intensity for all vocal tasks in with and without wearing either a surgical mask or a KN95 mask. In another study by Fiorella et al. (2021), wearing a surgical mask did not lead to a noteworthy decrease in speech intensity during the production of a sustained vowel. However, at the individual level, it was observed that 65% of the participants experienced a reduction in speech intensity while wearing a surgical mask, with 35% demonstrating an increase in speech intensity. The authors proposed that certain speakers might adopt strategies to modify their speech to overcome the filtration effects of wearing masks. In addition to the speakers' vocal efforts as a compensatory mechanism, mask types and individual variations can influence the results. For example, it has been reported that while surgical masks do not show significant changes in intensity, FFP2 masks do reduce the intensity of produced speech (Maryn et al., 2021). Following Fiorella et al. (2021) reporting interindividual variations in the effects of face masks on speech amplitude, we also inspected the intensity difference for individual speakers, and it turned out that the difference was higher in the mask condition for three of the participants (P2, P5, P9, see [supplementary material](#)). For one speaker (P1), the opposite scenario was found, with the remaining speakers showing no difference. Comparing our results with the findings of all the previous research, we also conclude that there is a range of responses among participants when it comes to the effect of masks on speech amplitude. Specifically, we attribute the increase in amplitude to speakers' vocal efforts to adjust their speech in order to enhance their intelligibility.

In this study, we also addressed the question of whether specific acoustic parameters, word duration and mean intensity difference, interact with orofacial gestures and what principles underlie this probable interaction. We established our possible assumptions on the basis of two hypotheses – namely, the trade-off hypothesis suggesting that speakers use gestures to enhance the perception of their message (Alibali et al., 2001; Bangerter, 2004; De Ruiter, 2006; De Ruiter et al., 2012; Melinger & Levelt, 2004) and the hand-in-hand hypothesis, which views gestures redundant with speech suggesting that speakers use gestures for their own benefit (Goldin-Meadow, 2009; Goldin-Meadow et al., 2001; Kita, 2000; So et al., 2009). Following the trade-off hypothesis, we would expect a compensation effect with more pronounced orofacial expressions compensating for the degraded acoustic signal in whispered speech mode to facilitate the listener's perception of the whispered speech. On the other hand, if the hand-in-hand hypothesis conformed

to our results, we would expect speakers to use orofacial gestures for their own internal purposes and in parallel with their speech. In other words, following this hypothesis, we would not expect any changes in facial expressions under even 'marked' communicative settings.

Our results reveal positive correlations between word duration and right eyebrow movement, lip aperture and eye opening, although we did not find a correlation between mean amplitude difference and orofacial expressions. This finding indicates that orofacial gestures are realised in parallel with longer acoustic duration, which in turn lends support to the hand-in-hand hypothesis. However, we do not jump to this conclusion, preferring to delineate another thorough picture on the parallel enhancement of acoustic information and orofacial gestures. Whispering is a speech mode with reduced intensity and an absence of F0 compared with normal speech. Our findings indicate that it is not solely gestures that respond to the distorted signal; rather, other acoustic parameters, such as duration, also exhibit significant sensitivity to changes in communication settings. Given that speakers have limited options for expressing intonation in whispered speech, whispering may force them to make more intense efforts to be understood, which can simultaneously exert pronounced orofacial gestures and longer acoustic duration. This can be predicted by Lindblom (1990), who claims that individuals adjust their level of speech signal based on the needs of their audience. When listeners need a higher amount of acoustic information, speakers tend to hyper-articulation. Conversely, when listeners can supplement the acoustic input with information from alternative sources, speakers tend to lessen their articulatory effort. In the case of whispering, the listener's requirements regarding how speakers convey intonation may be greater, leading speakers to employ more exaggerated or hyper-speech patterns.

The integration of mask conditions showed that eyebrow raising is higher when whispering with a face mask. As already established, masks may not only act as a barrier to specific acoustic signals, but also degrade specific visual signals by covering the lower face, which can cause a decrement in speech perception. In such a context, where part of the auditory and visual information is distorted, upper face gestures – in our case eyebrow raising – may be more expressive (Mheidly et al., 2020) to make up for the absence of visual cues produced by lips and those acoustic characteristics affected by masks as well. Degraded acoustic cues and lip invisibility are mutually compensated for by an enhancement of particular acoustic properties of speech that are not affected by a mask, such as word duration. Consequently, a cumulative effect will happen in which both gestures of the upper face and acoustics are strengthened in parallel. We expand upon previous research by suggesting a mutually beneficial and synergistic interaction between speech and gesture. Our data provide evidence that the interaction between speech signals and gesturing involves not only compensatory effects but also adjustments in response to signal distortion. When the speech signal is distorted, compensation occurs both in the acoustic domain and in more intensified orofacial expressions. Similarly, when visual cues are distorted, both facial expressions and duration are intensified. This interplay highlights the close relationship between speech signals and gesturing, demonstrating their interconnected nature. In fact, a similar result was obtained for German by Žygis and Fuchs (2023), who have shown that both whispered speech and invisibility conditions (under which speakers did not see each other at all) induced more intense orofacial expressions and longer word durations in parallel. Thus, this study broadens the spectrum of research.

Regarding the interpretation of our observations in terms of the ‘trade-off’ and ‘hand-in-hand’ hypotheses outlined above, it is worth noting that the initial concepts of both focus primarily on the frequency of referential gestures. However, subsequent research has expanded the scope of gestures to include orofacial expressions such as eyebrow and head movements. These gestures have been evaluated in relation to prosodic aspects such as the perception of focus, prominence or intonation. Several studies have demonstrated a connection between orofacial movements, including those of the eyebrows and head, and prosodic cues (Cruz et al., 2017). We propose that the evaluation of both hypotheses should encompass a broad range of gestures and speech types while considering their mutual influence. In our specific case, the combination of whispered speech and orofacial expressions provides an optimal scenario for examining potential compensatory effects.

Furthermore, when mentioning the intensity of gestures, we specifically refer to the measurable extent of movement rather than their frequency. Similarly, we carefully analyse the magnitude of acoustic parameters and establish correlations with gestures. While we recognise that referential gestures and prosodic gestures have inherent differences, we propose that an interaction between gesturing and speech can be investigated in both scenarios.

There also exists a distinct disparity between the original material employed for evaluating these two hypotheses and the material in this and other similar studies. Our study is limited to a controlled laboratory experiment, and a similar investigation in more natural settings is required for more comprehensive conclusions. Similar results should also be evaluated in spontaneous speech with more ecologically valid conditions as a possible research path. The scope of the present study is confined to examining speech production. It does not extend to evaluating how speakers perceive various acoustic and visual cues within different contexts. In order to explore the relative weight of orofacial expressions and acoustics, it is imperative to conduct complementary perceptual experiments examining the prosody of questions and statements. This study serves as an initial stage, aimed at acquiring essential information for use in future perceptual experiments.

In this investigation, we did not measure head movements. While we focussed on orofacial expressions to distinguish questions versus declaratives, the absence of head movements as a potential cue might have limited the comprehensiveness of our analysis. The decision to exclude head movements from our analysis was primarily based on the scope of our study, which aimed to investigate the role of orofacial expressions in distinguishing questions from statements. Additionally, given the complexities of analysing multiple non-verbal cues simultaneously, we chose to narrow our focus to orofacial expressions as a primary cue for differentiation. While orofacial expressions provided valuable insights into differentiating questions from statements, it is essential to acknowledge that head movements can also play a significant role in multimodal prosody. To build upon our findings, future research could explore the combined effects of orofacial expressions and head movements in distinguishing questions from statements.

We should also point out the limitations of the technique used for the detection and measurement of orofacial movements. We employed the OpenCV facial landmark detector, which, while generally reliable, encountered difficulties in accurately locating the landmarks on the eyebrows and eyes of two participants when they wore a face mask. Future studies in this area might consider improved techniques for more

robust detection of orofacial movements in situations where participants wear face coverings or accessories that could obscure facial features.

In summary, the results reported here showed the employment of orofacial gestures and acoustic cues to execute prosodic patterns in different communicative settings in Farsi. To the best of the author's knowledge, this is the first experimental study conducted for this language. Its results correspond to those obtained in a previous study on a typologically different language, namely German (Żygis & Fuchs, 2023). Future research scrutinising more languages will show whether the results are cross-linguistically valid.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/langcog.2024.21>.

Acknowledgements. We would like to thank Pillar Prieto, the editor of this special issue and three anonymous reviewers for their insightful comments. We also thank our participants for taking part in the experiment.

Funding statement. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, project "Audio-visual prosody of (semi-)whispered speech" ZY 117/4-1, PI: Marzena Żygis).

Competing interest. The authors declare none.

References

- Alibali, M. W., Heath, D. C., & Myers, H. J. (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.
- Ambrazaitis, G., & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, 95, 100–113.
- Arvaniti, A., Ladd, D. R., & Mennen, I. (2006). Phonetic effects of focus and "tonal crowding" in intonation: Evidence from Greek polar questions. *Speech Communication*, 48(6), 667–696.
- Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6), 415–419.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2020). *Package 'lme4'* (version 1), <https://cran.r-project.org/web/packages/lme4/lme4.pdf>.
- Bavelas, J., Gerwing, J., & Healing, S. (2014). Hand and facial gestures in conversational interaction. In T. M. Holtgraves (Ed.), *The Oxford handbook of language and social psychology* (pp. 111–130). Oxford University Press.
- Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.40. Retrieved 5 July 2018 from <http://www.praat.org/>.
- Borràs-Comes, J., Kaland, C., Prieto, P., & Swerts, M. (2014). Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior*, 38(1), 53–66. <https://doi.org/10.1007/s10919-013-0162-0>.
- Borràs-Comes, J., & Prieto, P. (2011). 'Seeing tunes.' The role of visual gestures in tune interpretation. *Laboratory Phonology*, 2(2), 355–380.
- Bould, E., & Morris, N. (2008). Role of motion signals in recognizing subtle facial expressions of emotion. *British Journal of Psychology*, 99(2), 167–189.
- Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles*, 32(1–2), 79–90.
- Buck, R., Miller, R. E., & Caul, W. F. (1974). Sex, personality, and physiological variables in the communication of affect via facial expression. *Journal of Personality and Social Psychology*, 30(4), 587–596.

- Buck, R. W., Savin, V. J., Miller, R. E., & Caul, W. F. (1972). Communication of affect through facial expressions in humans. *Journal of Personality and Social Psychology*, 23(3), 362–371.
- Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and Fo variations. *Proceeding of Fourth International Conference on Spoken Language Processing*, 4, 2175–2178. <https://doi.org/10.1109/ICSLP.1996.607235>.
- Cohn, M., Pycha, A., & Zellou, G. (2021). Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition*, 210, 104570.
- Corey, L., Mascola, J. R., Fauci, A. S., & Collins, F. S. (2020). A strategic approach to COVID-19 vaccine R&D. *Science*, 368(6494), 948–950.
- Corey, R. M., Jones, U., & Singer, A. C. (2020). Acoustic effects of medical, cloth, and transparent face masks on speech signals. *Journal of the Acoustical Society of America*, 148(4), 2371–2375.
- Crespo-Sendra, V., Kaland, C., Swerts, M., & Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, 47(1), 1–13.
- Cruz, M., Pejovic, J., Severino, C., Vigário, M., & Frota, S. (2022). Auditory and visual cues in face-masked infant-directed speech. In *Proceedings of the International Conference on Speech Prosody, Lisbon, Portugal* (pp. 639–643). <https://doi.org/10.21437/SpeechProsody.2022-130>.
- Cruz, M., Swerts, M., & Frota, S. (2015). Variation in tone and gesture within language. In *Proceedings of the 18th international congress of phonetic sciences, The Scottish Consortium for ICPHS 2015*. Glasgow, UK: the University of Glasgow. <http://hdl.handle.net/10451/25020>.
- Cruz, M., Swerts, M., & Frota, S. (2017). The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties. *Laboratory Phonology*, 8(1), 1–24. <https://doi.org/10.5334/labphon.110>.
- De Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate?. *Advances in Speech Language Pathology*, 8(2), 124–127.
- De Ruiter, J. P., Bangertler, 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(2), 232–248.
- Delis, I., Chen, C., Jack, R. E., Garrod, O. G., Panzeri, S., & Schyns, P. G. (2016). Space-by-time manifold representation of dynamic facial expressions for emotion categorization. *Journal of Vision*, 16(8), 14.
- Dimberg, U., & Lundquist, L. O. (1990). Gender differences in facial reactions to facial expressions. *Biological Psychology*, 30(2), 151–159.
- Dohen, M., & Loevenbruck, H. (2008). Audiovisual perception of prosodic contrastive focus in whispered French. *Journal of the Acoustical Society of America*, 123(5), 3460.
- Dohen, M., Loevenbruck, H., Cathiard, M. A., & Schwartz, J. L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, 44(1–4), 155–172.
- Ekman, P., Nelson, C. A., Horowitz, F. D., Spinrad, S. I., Sackeim, H. A., & Gur, R. C. (1980). Asymmetry in facial expression. *Science*, 209(4458), 833–836. <http://www.jstor.org/stable/1684659>.
- Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research: JSLHR*, 56(3), 850–864.
- Fan, X., Godin, K. W., & Hansen, J. H. (2011). Acoustic analysis of whispered speech for phoneme and speaker dependency. In *Twelfth annual conference of the international speech communication association* (pp. 181–184).
- Fiorella, M. L., Cavallaro, G., Di Nicola, V., & Quaranta, N. (2021). Voice differences when wearing and not wearing a surgical mask. *Journal of Voice*, 37(3), 467.e1–467.e7. <https://doi.org/10.1016/j.jvoice.2021.01.026>.
- Fonagy, J. (1969). Accent et intonation dans la parole chuchotée. *Phonetica*, 20(2–4), 177–192.
- Forni-Santos, L., & Osório, F. L. (2015). Influence of gender in the recognition of basic facial expressions: A critical literature review. *World Journal of Psychiatry*, 5(3), 342–351.
- Frota, S. (2002). Nuclear falls and rises in European Portuguese: a phonological analysis of declarative and question intonation. *Probus*, 14(1), 113–146.
- Frühholz, S., Trost, W., & Grandjean, D. (2016). Whispering—the hidden side of auditory communication. *NeuroImage*, 142, 602–612.
- Georgiou, G. P. (2022). Acoustic markers of vowels produced with different types of face masks. *Applied Acoustics*, 191, 108691. <https://doi.org/10.1016/j.apacoust.2022.108691>.
- Goldin, A., Weinstein, B., & Shiman, N. (2020). How do medical masks degrade speech perception. *Hearing Review*, 27(5), 8–9.

- Goldin-Meadow, S. (2009). How gesture promotes learning throughout childhood. *Child Development Perspectives*, 3(2), 106–111.
- Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, 12(6), 516–522.
- Guellai, B., Langus, A., & Nespors, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology*, 5, 700.
- Heeren, W. F. (2015a). Coding pitch differences in voiceless fricatives: Whispered relative to normal speech. *Journal of the Acoustical Society of America*, 138(6), 3427–3438.
- Heeren, W. F. (2015b). Vocalic correlates of pitch in whispered versus normal speech. *Journal of the Acoustical Society of America*, 138(6), 3800–3810.
- Heeren, W. F., & Lorenzi, C. (2014). Perception of prosody in normal and whispered French. *Journal of the Acoustical Society of America*, 135(4), 2026–2040.
- Heeren, W. F. L., & van Heuven, V. J. (2009). Perception and production of boundary tones in whispered Dutch. *Proceedings of Interspeech*, 2009, 2411–2414.
- Hellige, J. B. (1993). *Hemispheric asymmetry: What's right and what's left*. Harvard University Press.
- Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, 17(2), 142–157. <https://doi.org/10.1177/1088868312472607>.
- Higashikawa, M., & Minifie, F. D. (1999). Acoustical-perceptual correlates of “whisper pitch” in synthetically generated vowels. *Journal of Speech, Language, and Hearing Research*, 42(3), 583–591.
- Higashikawa, M., Nakai, K., Sakakura, A., and Takahashi, H. (1996). Perceived pitch of whispered vowels—Relationship with formant frequencies: A preliminary study. *Journal of Voice*, 10(2), 155–158.
- Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639–652.
- Hömke, P. (2019). *The face-to-face communication. Signals of understanding and non-understanding*. [Doctoral dissertation]. Max Planck Institute, Nijmegen.
- Hömke, P., Levinson, S. C., & Holler, J. (2022). *Eyebrow movements as signals of communicative problems in human face-to-face interaction*. PsyArXiv. <https://doi.org/10.31234/osf.io/3jnmnt>.
- House, D. (2002). Intonational and visual cues in the perception of interrogative mode in Swedish. In *Proceedings of 7th international conference on spoken language processing, ICSLP2002 - Interspeech 2002, Denver, Colorado, USA* (pp. 1957–1960).
- Joshi, A., Procter, T., & Kulesz, P. A. (2021). COVID-19: Acoustic measures of voice in individuals wearing different face masks. *Journal of Voice*, 37, 971.e1–971.e8. <https://doi.org/10.1016/j.jvoice.2021.06.015>.
- Jović, S. T., & Šarić, Z. (2008). Acoustic analysis of consonants in whispered speech. *Journal of Voice*, 22(3), 263–274.
- Kallail, K. J., & Emanuel, F. W. (1984). “An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects,” *Journal of Phonetics*, 12(2), 175–186.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10, 1755–1758.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.
- Knowles, T., & Badh, G. (2022). The impact of face masks on spectral acoustics of speech: Effect of clear and loud speech styles. *Journal of the Acoustical Society of America*, 151(5), 3359–3368.
- Kong, Y. Y., & Zeng, F. G. (2006). Temporal and spectral cues in Mandarin tone recognition. *Journal of the Acoustical Society of America*, 120(5), 2830–2840.
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.
- Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process. *Language and Gesture*, 2(261), 261–283.
- Krumhuber, E., Manstead, A. S. R., & Kappas, A. (2007). Temporal aspects of facial displays in person and expression perception: The effects of smile dynamics, head-tilt, and gender. *Journal of Nonverbal Behavior*, 31(1), 39–56.
- Length, R (2019). Emmeans: Estimated Marginal Means, aka Least-Squares Means. <https://cran.r-project.org/web/packages/emmeans/index.html>.

- Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130302.
- Lin, Y., Cheng, L., Wang, Q., & Xu, W. (2021). Effects of medical masks on voice assessment during the COVID-19 pandemic. *Journal of Voice*, 37, 802.e25–802.e29. <https://doi.org/10.1016/j.jvoice.2021.04.028>.
- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In P. MacNeilage & J. L. Davis (Eds.), *Speech production and speech modelling* (pp. 403–439). Springer Netherlands.
- Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralised. *Language and Speech*, 47(2), 109–138.
- Magee, M., Lewis, C., Noffs, G., Reece, H., Chan, J. C., Zaga, C. J., Paynter, C., Birchall O., Azocar S. R., Ediriweera A., Kenyon K., Caverl  M. W., Schultz, B. G., & Vogel, A. P. (2020). Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols. *Journal of the Acoustical Society of America*, 148(6), 3562–3568. <https://doi.org/10.1121/10.0002873>.
- Maryn, Y., Wuys, F. L., & Zarowski, A. (2021). Are acoustic markers of voice and speech signals affected by nose-and-mouth-covering respiratory protective masks?. *Journal of Voice*, 35(2), 468.e1–468.e12.
- Massaro, D., & Cohen, M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology Human Perception & Performance*, 9(5), 753.
- McDuff, D., Kodra, E., Kaliouby, R. E., & LaFrance, M. (2017). A large-scale analysis of sex differences in facial expressions. *PLoS One*, 12(4), e0173942.
- McKenna, V. S., Kendall, C. L., Patel, T. H., Howell, R. J., & Gustin, R. L. (2022). Impact of face masks on speech acoustics and vocal effort in healthcare professionals. *Laryngoscope*, 132(2), 391–397. <https://doi.org/10.1002/lary.29763>.
- McKenna, V. S., Patel, T. H., Kendall, C. L., Howell, R. J., & Gustin, R. L. (2021). Voice acoustics and vocal effort in mask-wearing healthcare professionals: A comparison pre- and post-workday. *Journal of Voice*, 37, 802.e15–802.e23. <https://doi.org/10.1016/j.jvoice.2021.04.016>.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
- McNeill, D., Quek, F., McCullough, K. E., Duncan, S., Furuyama, N., Bryll, R., Ma, X. F., & Ansari, R. (2001). Catchments, prosody and discourse. *Gesture*, 1(1), 9–33.
- Melinger, A., & Levelt, W. J. (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4(2), 119–141.
- Mendoza-Denton, N., & Jannedy, S. (2011). Semiotic layering through gesture and intonation: A case study of complementary and supplementary multimodality in political speech. *Journal of English Linguistics*, 39(3), 265–299.
- Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America*, 29(1), 104–106.
- Mheidly, N., Fares, M. Y., Zalzale, H., & Fares, J. (2020). Effect of face masks on interpersonal communication during the COVID-19 pandemic. *Frontiers in Public Health*, 8, 582191. <https://doi.org/10.3389/fpubh.2020.582191>.
- Miller, J. D. (1961). Word tone recognition in Vietnamese whispered speech. *Word*, 17(1), 11–15.
- Miranda, L., Gomes da Silva, C., Moraes, J. A. & Rilliard, A. (2020). Visual and auditory cues of assertions and questions in Brazilian Portuguese and Mexican Spanish: A comparative study. *Journal of Speech Sciences*, 9, 73–92.
- Miranda, L., Swerts, M., Moraes, J., & Rilliard, A. (2021). The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo Questions. *Language and Speech*, 64(1), 3–23. <https://doi.org/10.1177/0023830919898886>.
- Miranda, L. S., Moraes, J. A., & Rilliard, A. (2019). Audiovisual perception of wh-questions and wh-exclamations in Brazilian Portuguese. In *Proceedings of 19th international congress of phonetic sciences, Melbourne, Australia* (pp. 2941–2945).
- Nguyen, D. D., McCabe, P., Thomas, D., Purcell, A., Doble, M., Novakovic, D., Chacon, A., & Madill, C. (2021). Acoustic voice characteristics with and without wearing a facemask. *Scientific Reports*, 11(1), 5651. <https://doi.org/10.1038/s41598-021-85130-8>.
- Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11(8), 1017.
- OpenCV. (2015). *Open Source Computer Vision Library*.

- Palmiero, A. J., Symons, D., Morgan III, J. W., & Shaffer, R. E. (2016). Speech intelligibility assessment of protective facemasks and air-purifying respirators. *Journal of Occupational and Environmental Hygiene*, 13(12), 960–968.
- Pörschmann, C., Lübeck, T., & Arend, J. M. (2020). Impact of face masks on voice radiation. *Journal of the Acoustical Society of America*, 148(6), 3663–3670.
- Pouw, J., & Trujillo, W. L. (2021 November 18). *Multimodal Face-Prosody Analysis Using OpenFace and Parselmouth*. <https://github.com/WimPouw/EnvisionBootcamp2021/tree/main/Python/MediaBodyTracking>.
- Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49, 41–54.
- Rahne, T., Fröhlich, L., Plontke, S., & Wagner, L. (2021). Influence of surgical and N95 face masks on speech perception and listening effort in noise. *PLoS One*, 16(7), e0253874. <https://doi.org/10.1371/journal.pone.0253874>.
- Rossano, F. (2010). Questioning and responding in Italian. *Journal of Pragmatics*, 42(10), 2756–2771.
- RStudio Team (2022). *RStudio: integrated development for R*. Rstudio, Inc. (Version 4.2.2) [Computer software].
- Sadat-Tehrani, N. (2011). The intonation patterns of interrogatives in Persian. *Linguistic Discovery*, 9(1), 105–136.
- Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In *Proceedings of the IEEE international conference on computer vision workshops, Sydney, Australia* (pp. 397–403).
- Sardhaei, N., Żygis, M., and Sharifzadeh, H. (2022). How do our eyebrows respond to masks and whispering? The case of Persian. *Proceedings of Interspeech, 2022, 2023–2027*. <https://doi.org/10.21437/Interspeech.2022-10867>.
- Sato, W., & Yoshikawa, S. (2004). The dynamic aspects of emotional facial expressions. *Cognition and Emotion*, 18(5), 701–710.
- Sato, W., & Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition*, 104(1), 1–18.
- Schmidt, K. L., Bhattacharya, S., & Denlinger, R. (2009). Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. *Journal of Nonverbal Behavior*, 33(1), 35–45.
- Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, 9, 1514.
- Sinagra, C., & Wiener, S. (2022). The perception of intonational and emotional speech prosody produced with and without a face mask: An exploratory individual differences study. *Cognitive Research: Principles and Implications*, 7(1), 89.
- So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33(1), 115–125.
- Sowden, S., Schuster, B. A., Keating, C. T., Fraser, D. S., & Cook, J. L. (2021). The role of movement kinematics in facial emotion expression production and recognition. *Emotion*, 21(5), 1041–1061.
- Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech*, 46(1), 1–22.
- Tao, F., & Busso, C. (2014). Lipreading approach for isolated digits recognition under whisper and neutral speech. In *Proceedings of 15th annual conference of the international speech communication association, Interspeech 2014, Singapore* (pp. 1154–1158).
- Tartter, V. C. (1989). What's in a whisper?. *Journal of the Acoustical Society of America*, 86(5), 1678–1683.
- Torreira, F., & Valtersson, E. (2015). Phonetic and visual cues to questionhood in French conversation. *Phonetica*, 72(1), 20–42.
- Van der Sluis, I., & Kraemer, E. (2007). Generating multimodal references. *Discourse Processes*, 44(3), 145–174.
- Wallbott, H. G. (1988). Big girls don't frown, big boys don't cry? Gender differences of professional actors in communicating emotion via facial expression. *Journal of Nonverbal Behavior*, 12(2), 98–106.
- Winter, B. (2020). *Statistics for Linguists. An Introduction using R*. Routledge.
- Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30(3), 555–568.

- Zeng, X. L., Martin, P., & Boulakia, G. (2004). Tones and intonation in declarative and interrogative sentences in Mandarin. In *Proceedings of the international symposium on tonal aspects of languages with emphasis on tone languages, Beijing, China* (pp. 235–238).
- Żygis, M., & Fuchs, S. (2019). How Prosody, speech mode and speaker visibility influence lip aperture. In *Proceedings of the 19th international congress of phonetic sciences* (pp. 230–234).
- Żygis, M., & Fuchs, S. (2023). Communicative constraints affect oro-facial gestures and acoustics: Whispered vs normal speech. *Journal of the Acoustical Society of America*, 153(1), 613–626.
- Żygis, M., Pape, D., Koenig, L. L., Jaskuła, M., & Jesus, L. M. (2017). Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes. *Journal of Phonetics*, 63, 53–74.

Cite this article: Mahdinazhad Sardhaei, N., Żygis, M., & Sharifzadeh, H. (2024). Facial expressions in different communication settings: A case of whispering and speaking with a face mask in Farsi, *Language and Cognition*, 1–35. <https://doi.org/10.1017/langcog.2024.21>