Interspeech 2022
18-22 September 2022, Incheon, Korea

How do our eyebrows respond to masks and whispering? The case of Persian

Nasim Mahdinazhad Sardhaei[1], *Marzena Żygis*[1,3], *Hamid Sharifzadeh*[2]

[1] Leibniz-Centre General Linguistics (Leibniz-ZAS), Berlin
[2] Unitec Institute of Technology, Auckland, New Zealand
[3] Humboldt Universität zu Berlin, Berlin

sardhaei@leibniz-zas.de, zygis@leibniz-zas.de, hsharifzadeh@unitec.ac.nz

Abstract

Whispering is one of the mechanisms of human communication to convey linguistic information. Due to the lack of vocal fold vibration, whispering acoustically differs from the voiced speech in the absence of fundamental frequency which is one of the main prosodic correlates of intonation. This study addresses the importance of facial cues with respect to acoustic cues of intonation. Specifically, we aim to probe how eyebrow velocity and furrowing change when people whisper and wear face masks, also, when they are supposed to produce a prosodic modulation as it is the case in polar questions with rising intonation. To this end, we run an experiment with 10 Persian speakers. The results show the greater mean speed when speakers whisper indicating a compensation effect for the lack of F0 in whispering. We also found a more pronounced movement of both eyebrows when the speakers wear a mask. Finally, our results reveal greater eyebrow motions in questions suggesting the question is a more marked utterance type than a statement. No significant effect of eyebrow furrowing was found. However, eyebrow movements were positively correlated with the eyebrow widening suggesting a mutual link between these two movement types.

Index Terms: audiovisual prosody, facial gesture, respiratory protection mask, sentence type

1. Introduction

Human communication is essentially a multimodal form of signaling in which vocal and visual modes of communication are tightly interwoven. People integrate auditory and visual information produced by face and body to communicate intended meanings [1]. One of the core questions on the association between visual and spoken components of language is how various visual movements accompanying the speech contribute to the way the speech is uttered i.e., speech prosody?

Traditionally, speech prosody has been viewed via purely auditory/acoustic channels, but more recently there has been a growing awareness that the production and perception of speech including prosody are multimodal. A substantial number of studies have demonstrated the prominent contribution of facial expressions and other forms of visual correlates to the auditory properties of prosody [2], [3], [4] [5], [6]. For instance, an early study by [7] reported the dominance of rapid head movements over speech sounds in applying high intensities in conversation, or there have been claims that eyebrow movements are correlated with acoustic features of prosody, such as fundamental frequency and amplitude [8], [9]. Similarly, it has been noted in [10], [11], [12], [13] that facial gestures such as eyebrow movements map onto the intonation of questioning and responding. [11] examining facial gesture expressions in Catalan, showed the use of eyebrow lowering for expressing the echo questions. In contrast, Dutch polar questions were accompanied by raised eyebrows [14]. However, all these studies have tended to examine the relationship between visual and acoustic correlates of prosody in voiced speech. Relevant literature has merely acknowledged the role of facial gestures such as eyebrow movements in other vocal modes of speech such as whispering.

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. The question arising here is whether the deployment of facial gestures differs in voiced vs. whispered speech. One possible hypothesis to answer this question can be explained in terms of trading relations. Speakers to improve their intelligibility and compensate for the absence of acoustic cues while whispering, may resort to the visual channel of speech and exaggerate facial movements to express their intention. [15] in their study concluded gestures carry most of the communicative burden in the presence of background noise to compensate for the disruption of the auditory signal. According to the trade-off hypothesis, different modalities can be used to compensate for another based on the requirements of situational constraints [16]. In this paper, we attempt to explore whether eyebrow movements also enter a trade-off relation with speech signal when the latter lacks fundamental frequency [17], i.e., whether eyebrows are more pronounced when whispering, contributing more intensively to the understanding of speech?

Furthermore, we aim to find out if eyebrow movements are larger in questions than in statements. Except for a scant number of studies such as those conducted by [18], [19] who noticed the compensatory effect of facial gestures in perception enhancement of whispered speech as well as the study by [20] which found the involvement of higher eyebrow raises in whispered questions than in statements, questions of this type have been sought far scarcely in the literature.

Finally, we extend the spectrum of conditions in our study by investigating whether the relation between eyebrow movements and speech is in the function of mouth covering. Despite its high capability to reduce the risk of infection, wearing protective masks from the beginning of the COVID-19 pandemic interferes with spoken communication and leads to acoustic attenuation [21], [22]. The study by [23] reveals that wearing masks significantly influences the acoustic markers relevant to clinical speech including variations in fundamental frequency. Based on this, we formulate the hypothesis that when a mask is used, mutual compensation

10.21437/Interspeech.2022-10867

effect across modalities and speech modes will be stronger in both cases of degraded whisper and voiced speech, consequently impacting the motion of eyebrows in two speech modes in comparison to the condition when the speakers do not wear a mask.

With the goal of filling these gaps, this study aims to scrutinize the nature of eyebrow motion including the speed of movement and furrowing/widening focusing on three factors:

(a) the speech mode (whispered vs. normal speech)
(b) the mouth covering of speakers (with a mask vs. without a mask condition)
(c) the pragmatic function of a message related to prosody (polar questions with rising F0 vs. statements with non-rising F0).

2. Methods

2.1. Participants

A total of 10 native speakers of Persian completed experimental video recording sessions. The participants were 5 females and 5 males within the age group of 20 to 35 (mean: 30.6, SD: 5.03). They reported no hearing or speech impairments. The participants filled in a short demographic questionnaire and provided written informed consent, acknowledging their participation in the experiment and for demographic questionnaire use.

2.2. Stimuli

10 pairs of Persian statements and questions were created based on the stress pattern of the Persian language. The content of questions and statements was identical differing only in the punctuation, i.e., statements ended in a full stop and questions in a question mark. See the example below:

Question: Diruz Nina goft Kebab?
< Yesterday, Nina said Kebab? >
Statement: Diruz Nina goft Kebab.
< Yesterday, Nina said Kebab. >

The word order in a simple Persian sentence is SOV (*Subject, Object,* and *verb*), i.e., the verb normally comes at the end of a sentence. But we preferred to have a noun at the final position of the sentence since Persian verbs carry aspect, tense, and agreement, and they are usually accompanied by inflectional suffixes and prefixes which affect the regular stress pattern of the verbs. Thus, we decided to create quote sentences such as *she said, "target word"* in which the final word was a quoted noun. As the result, in this study, each sentence consisted of 4 words with the order of *adverb of time + subject + transitive verb + object*, with the *object* as the target word in the final position of each sentence. The target words were all bisyllabic content words with the stress falling on the second syllable. Not for the purpose of the current study, but as a part of a larger research project, the second syllable of all the controlled words had a CVC structure starting with a bilabial stop /p/, /b/ or /m/ and followed by unrounded vowels /a/, /ɛ/ or /ɪ/ such as, for instance, [keʃˈmiɾ] = "cashmere", and [dæmˈbel] = "dumbbells".

2.3. Data collection

The recordings were obtained in a soundproof studio in Tabriz, Iran. Participants were seated with their heads at the center of a frame with a green solid uniform background, positioned 1 meter from a tripod-mounted video camcorder (Sony Alpha a6400 Mirrorless Camera with 16-50mm 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 on the camcorder, and displayed the stimuli so that the participants could simultaneously look at the lens of the camera and read the stimuli on the screen without eye or head tilt. In addition to the video recordings, auditory data was synchronously recorded 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, digitized mono. In order to prevent head movements and ensure the fixed distance to the lens of the camera, we emphasized participants to keep their heads still. The participants were asked to practice ensuring they understood the procedure. Once, the participants felt ready, they uttered 4 lists each consisting of 20 randomized sentences. The presentation of the stimuli within these lists was randomized to prevent order effects. Speakers' facial movements were recorded in two conditions, wherein they produced pairs of questions and statements in voiced and whispered modes either wearing a disposable surgical mask on their faces or without the mask. In total, 80 sentences (20 sentences * 2 conditions * 2 speech modes) were recorded for each speaker.

2.4. Data preparation

The acoustic onset and offset of the final content word in each sentence were labeled using Praat (version 6.0.28) [24]. The extracted time points were passed to a customized python script to crop each video recording into the given time points and save out the output video files for each of the content words produced per speaker.

2.5. Video processing

Pursuing our research goal to quantify the eyebrow movements, we used the following pipeline in our analyses. A feature combination of two python libraries was utilized to achieve an accurate facial landmark detection in the recordings: OpenCV [25], a library of python binding, and the cross-platform Dlib facial landmark detector [26] containing a pretrained model. By means of these two open-source tools, videos were iterated frame by frame at 29. 969 frames per second (FPS). In every video frame, the face region was identified and 68 2D landmarks were mapped to distinctive units of the face. Then, the estimates of x, y coordinates for each landmark were extracted in pixels. Figure 1 illustrates the indexes of the 68 coordinates on the face of a participant. The subject's written consent was obtained to use his photograph for publication.

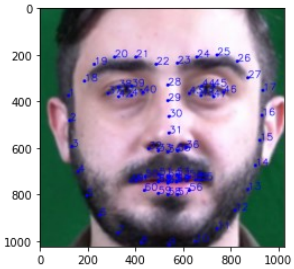

Figure 1: *68 Facial Landmarks Tracked by OpenCV.*

As an important requirement, the input videos (i.e., a collection of successive frames) needed to be normalized to minimize the probable unexpected head movements and scale

orientation of the face in each frame. To this end, we used an affine transformation by which the distance ratio between the center of each eye as well as the angle between the eye centroids was calculated. The midpoint between the left and right eyes, atop of nose, was considered the reference position serving as the (x, y) coordinate in which the face was rotated around so that the eyes lay along the same y coordinates. With this method, a canonical representation of the face was obtained. Once the x, y positions of facial landmarks were derived, the Euclidean distance *(D)* between the key points corresponding to each type of eyebrow movement was calculated for each successive frame as:

$$(D) = \sqrt{((x_1 - x_2)^{\wedge 2} + (y_1 - y_2)^{\wedge 2})} \quad (1)$$

We are interested in the vertical eyebrow movement to compute the movement velocity. Thus, we took the three points in the middle of eyebrows and calculated distance *D1* between mean of these central key points for each eyebrow (mean of landmarks 19, 20, and 21 for the right eyebrow, and mean of 24, 25, and 26 for the left eyebrow) and the key point in the middle of the nose (landmark 28) to obtain the vertical position of eyebrows (Figure 2).

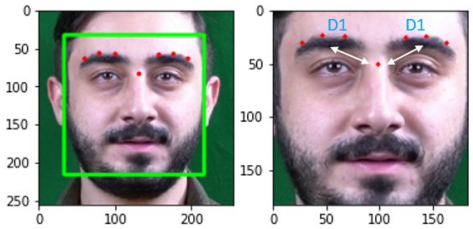

Figure 2: *Key points for the measurement of D1.*

The distance D2 between the two landmarks 22 on the right eyebrow and 23 on the left eyebrow was also measured to obtain the position of eyebrows when furrowed or widened (Figure 3).

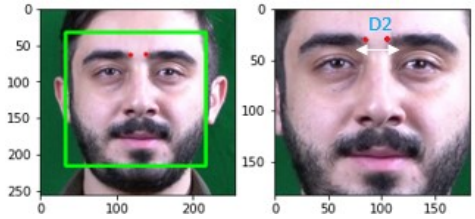

Figure 3: *Key points for the measurement of D2.*

In this study, speed was defined as the average frame to frame change in distance *(D)* between the target key points for each eyebrow across each video. More precisely, speed was calculated as the average summation of differentials (*Vn*) for each face action vector representing the absolute mean speed of movement within the given recording window (pixel/frame):

$$V(n)=D(n)–D(n\text{-}1) \quad (2)$$

where *n* is the number of each frame per video

$$Mean\ Speed = V_{(n1)} + V_{(n2)} + V_{(n3)} + \ldots\ / fnum \quad (3)$$

where *fnum* is the total number of frames per video.

Finally, the resulting speed vectors were low pass filtered by submitting them to a Savitzky-Golay filter [27], parameterized with a window size of seven samples and a third-degree polynomial.

2.6. Statistics

The statistical analysis was conducted in R 3.4.2 (R Core Team 2017) [28] by using package lme4 [29]. By means of linear mixed models, we investigated the effects of SPEECH MODE [normal, whispered], MASK CONDITION [with mask, without mask], and SENTENCE TYPE [question, statement] on the following dependent variables: LEFT EYEBROW MOVEMENT, RIGHT EYEBROW MOVEMENT, and EYEBROW FURROWING. We also included the interaction of SPEECH MODE and MASK CONDITION. Finally, a random structure was added to the models: ITEM and PARTICIPANT as random intercepts, and random by-participant and by-item slopes for the SPEECH MODE, MASK CONDITION, and SENTENCE TYPE. Due to non-convergence issues of the models, some random slopes were removed after examining their correlations. Apart from linear mixed models, we also calculated correlations between left and right eyebrow movements on the one hand and eyebrow movements with eyebrow furrowing/widening on the other.

3. Results

To probe the coordination between both eyebrows, we computed Pearson's correlation between left and right eyebrow movements. The results indicate statistically significant relationship between the movements of both eyebrows ($r=0.819$, $t(742) = 38.841$, $p<001$), see Figure 4.

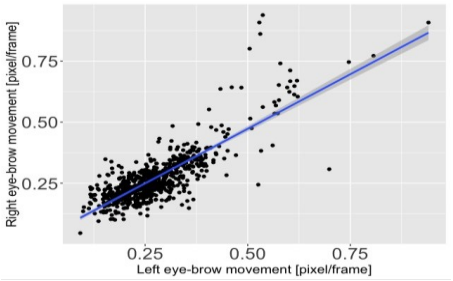

Figure 4*: The relationship between the left and right eyebrow mean speed*

Moreover, we searched for the relationship between the speed of vertical eyebrow and horizontal type of eyebrow movement. The results reveal a correlation between the two types of movements for the right eyebrow ($r=0.545$ ($t(741) = 17.717$, $p<001$). Similarly, for the left eyebrow, a significant correlation of 0.5 was found between the two movement types ($t(741) = 15.707$, $p<001$). The results are demonstrated in Figure 5.

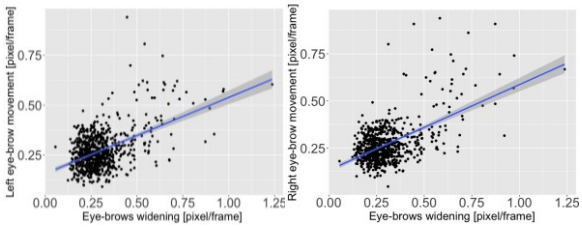

Figure 5*: The relationship between two movements*

The results of linear mixed model concerning the change of the left eyebrow reveal a significant influence of speech mode: the speed of movement is higher in whispered than normal speech ($\beta=0.047$, $t=5.29$, $p<.001$). It is also quicker when people wear masks as opposed to the condition without masks ($\beta=0.086$, $t=3.46$, $p<.01$) and higher when they produce questions as opposed to statements ($\beta=0.035$, $t=3.31$, $p<.01$).

The difference in the left eyebrow change between normal and whispered speech is also higher when people wear masks as opposed to their behavior without masks (SPEECH MODE*MASK CONDITION, β=-0.021, t=-2.21, p<.05). This is illustrated in Figure 6.

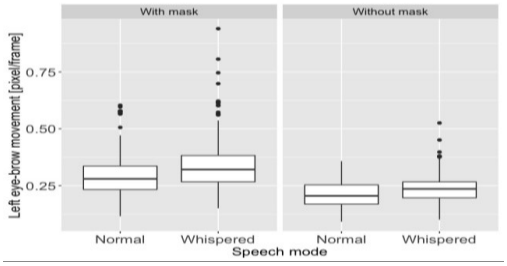

Figure 6: *Left eye-brow movement in speech modes*

Regarding the change of the right eyebrow, it is higher when speakers whisper (β=0.049, t=5.74, p<.001), wear a mask (β=0.091, t=10.46, p<.001), and produce questions (β=0.029, t=3.12, p<.05). In contrast to the change of the left eyebrow, the interaction of SPEECH MODE*MASK CONDITION was not significant, see Figure 7.

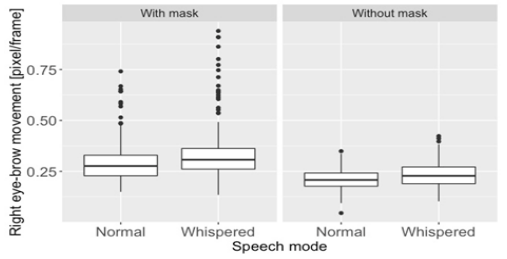

Figure 7*: Right eye-brow movement in speech modes*

Neither of the variables exerted a significant effect for the speed of eyebrow furrowing/widening (D2), see Figure 8.

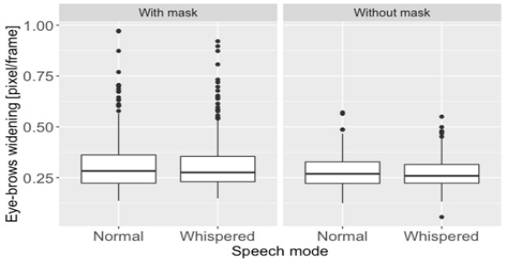

Figure 8*: Eyebrow widening* (D2) *in speech modes*

4. Discussion

In this study, we extend the repertoire of facial expressions by examining the role of eyebrow movements as a visual cue in production of whispered speech. Past work has largely examined the role of eyebrows movements in emotional expression or sign languages, and eyebrows have received little attention in the audiovisual research domain. Our core question to investigate was the potential interactions between degraded acoustic signals in whispering and eyebrow motions. Given the fact that different muscles drive eyebrow movements and due to the functional differences between the brain hemispheres during communication [30], we examined each eyebrow separately and found different effects emerging for the movement of each eyebrow in various conditions. First, the results show the speed of eyebrow movement is greater in whispered as opposed to voiced speech. Second, the velocity of eyebrows is higher in questions than statements.

In line with the earlier studies of [15] and [20], our findings support the trade-off hypothesis by confirming the compensation effects between the speech mode and the speed of eyebrow movement. The lack of F0 is compensated by a greater velocity of eyebrows which may enhance visual cues for the speech perception of the interlocutors. Moreover, it seems faster movements of eyebrows can enter a trade-off relationship with the acoustic signal to produce the difference between questions and statements. However, further acoustic analysis and specially perception experiments are needed to underpin our findings. Also, the coordination between the motion of eyebrow and other facial gestures such as lip aperture need to be examined for more comprehensive conclusions of the complex compensatory multimodal interaction.

We also made an effort to enlarge the spectrum of research on audiovisual properties of speech by an analysis of additional face mask condition. The use of face masks hinders the contribution of the middle and lower face in the expression of the message. As a result, the role of the upper face in drawing the attention of the listener increases in significance [31]. Establishing our hypothesis based on this assumption, we probed the interaction between eyebrow movement and speech mode as a function of face mask use. The outcomes indicate a trading relation between lower face covering and speech mode. The invisibility of the middle and lower face is made up by a greater velocity of eyebrows. According to [31], protective face masks cause a great decrement in speech perception and lead to the use of higher pitched voice. Taking this hindering effect of the mask into consideration and based on the outcome of our experiment, we conclude that since whispered speech lacks the fundamental frequency, which is the physical correlate of pitch, speakers may resort to the faster movement of their eyebrows. This can be interpreted as a coping strategy to compensate for the absence of pitch when whispering from behind the protective masks and facilitate delivering the message.

As far as the speed of eyebrow furrowing is concerned, no significant movement was detected in any of the experimental conditions, but we found a positive linear correlation between eyebrow velocity (D1) and furrowing/widening (D2). This positive correlation between horizontal and vertical movement can be explained in terms of the symmetrical behavior of eyebrows. If one of the eyebrows moves vertically higher, then the horizontal distance between two points can be expected to be longer too.

5. Conclusion

This study demonstrates the compensatory role of eyebrow movement in the absence of fundamental frequency in whispered speech. Both eyebrows also showed a greater mean speed when speakers used a face mask. In the prosodic condition where questions and statements were produced; a larger velocity of eyebrows was observed in the former utterance type. Contrarily, we did not find a significant effect of eyebrow furrowing in any of the conditions.

6. Acknowledgements

This work was supported by a DFG (Deutsche Forschungsgemeinschaft) (“Audio-visual prosody of (semi-) whispered speech” ZY 117/4-1). We would like to thank all the participants for their participation in our experiment.

7. References

[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 26, no. 4, pp.746–748, Dec. 1976.

[2] D. McNeill, F. Quek, K. E. McCullough, S. D. Duncan, N. Furuyama, R. Bryll, and R. Ansari, “Catchments, prosody and discourse,” *Gesture*, vol. 1, no. 1, pp. 9-33, June. 2001.

[3] E. J. Krahmer and M. Swerts, “The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception,” *Journal of Memory and Languag*e, vol. 57, no. 3, pp. 396–414, Oct. 2007.

[4] E. Krahmer and M. Swerts, “Audiovisual prosody—introduction to the special issue,” *Language and speech*, vol. 52, no. 2-3, pp. 129-133, June. 2009.

[5] N. Mendoza-Denton and S. Jannedy, “Semiotic layering through gesture and intonation: a case study of complementary and supplementary multimodality in political speech,” *Journal of English Linguistics*, vol. 39, no. 3, pp. 265–299, June. 2011.

[6] B. Guellaï, A. Langus, and M. Nespor, “Prosody in the hands of the speaker,” *Frontiers in Psychology*, vol. 5, no. 700, pp. 1–8, July. 2014.

[7] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, “Head movement correlates of juncture and stress at sentence level,” *Language and Speech*, vol. 26, no. 2, pp. 117–129, Apr. 1983.

[8] C. Cave, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, “About the relationship between eyebrow movements and F0 variations,” in *Proceedings of the International Conference on Spoken Language Processing (ICSLP),* Philadelphia, PA, Oct. 1996, pp. 2175-2179.

[9] I. Guaitella, S. Santi, B. Lagrue, and C. Cave´, “Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation,” *Language and Speech*, vol. 52, no. 2, pp. 207–222, June. 2009.

[10] D. House, “Intonation and visual cues in the perception of interrogative mode in Swedish,” in *Proceedings INTERSPEECH 2002 – 7th International Conference on Spoken Language Processing*, Colorado, USA, Sept. 2002, pp. 1957-1960.

[11] J. Borràs-Comes and P. Prieto, “‘Seeing tunes.’ The role of visual gestures in tune interpretation,” *Laboratory Phonology*, vol. 2, no. 2, pp. 355–380, Oct. 2011.

[12] L. S. Miranda, J. Moraes, and A. Rilliard, “Audiovisual perception of wh-questions and wh-exclamations in Brazilian Portuguese,” in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, Aug. 2019, pp. 2941–2945.

[13] M. Cruz, M. Swerts, and S. Frota, “The role of intonation and visual cues in the perception of sentence types: Evidence from European Portuguese varieties,” *Journal of the Association for Laboratory Phonology*, vol. 8, no. 1, pp. 1–24, Sept. 2017.

[14] J. Borràs-Comes, C. Kaland, P. Prieto, and M. Swerts, “Audiovisual correlates of Interrogativity: a comparative analysis of Catalan and Dutch,” *Journal of Nonverbal. Behavior,* vol. 38, no. 1, pp. 53–66, March. 2014.

[15] J. Trujillo, A. Özyürek, J. Holler, and L. Drijvers, “Speakers exhibit a multimodal Lombard effect in noise,” *Scientific Reports, vol.* 11, no. 16721, Aug. 2021.

[16] M. Żygis and S. Fuchs, “How prosody, speech mode and speaker visibility influence lip aperture,” in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, Aug. 2019, pp. 230 –234.

[17] G. A. Miller and P. Nicely, “An Analysis of Perceptual Confusions among some English Consonants,” *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338-352, May. 1955.

[18] M. Dohen and H. Loevenbruck, “Interaction of audition and vision for the perception of prosodic contrastive focus,” *Language and Speech*, vol. 52, no. 2–3, pp. 177-206, June. 2009.

[19] N. Mendoza-Denton and S. Jannedy, “Semiotic layering through gesture and intonation: a case study of complementary and supplementary multimodality in political speech,” *Journal of English Linguistics*, vol. 39, no. 3, pp. 265–299, June. 2011.

[20] M. Żygis, S. Fuchs, and K. Stoltmann, “Orofacial expressions in German questions and statements in voiced and whispered speech,” *Journal of Multimodal Communication Studies*, vol. 4, no. 1-2, pp. 87-92, Aug. 2017.

[21] C. Llamas, P. Harrison, D. Donnelly, and D. Watt, “Effects of different types of face coverings on speech acoustics and intelligibility,” York Papers in Linguistics Series 2, no. 9, pp. 80–104, 2008.

[22] C. Porschmann, T. J. Lubeck, and M. Arend, “Impact of face masks on voice radiation,” *Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 3663–3670, Dec. 2020.

[23] Y. Maryn, F. L. Wuyts, and A. J. Zarowski, “Are Acoustic Markers of Voice and Speech Signals Affected by Nose-and-Mouth-Covering Respiratory Protective Masks?,” *Journal of Voice*, Feb. 2021.

[24] P. Boersma and D. Weenink. *PRAAT: Doing phonetics by computer (Version 6.0.28)*. (2017).

[25] OpenCV. *Open-Source Computer Vision Library*. (2015).

[26] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machin Learning Research*, vol. 10, no. 60, pp.1755–1758, July. 2009.

[27] Savitzky, A.; Golay, M.J.E, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–39, July. 1964.

[28] R Core Team (Vienna, Austria: R Foundation for Statistical Computing). *‘R: A language and environment for statistical computing*. (2017).

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker. *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–17*. (2018).

[30] C. Cavé, Christian, I. Guaitella, R. Bertrand, S. Santi, and F. Harlay, "About the relationship between eyebrow movements and Fo variations," in *Proceeding of Fourth International Conference on Spoken Language Processing*. vol. 4, pp. 2175–2178, Oct. 1996.

[31] N. Mheidly, M. Y. Fares, H. Zalzale, and J. Fares, “Effect of face masks on interpersonal communication during the COVID-19 pandemic,” *Frontiers in Public Health*, vol. 8, no. 582191, pp. 1-6, Dec.2020.