HOW PROSODY, SPEECH MODE AND SPEAKER VISIBILITY INFLUENCE LIP APERTURE

Marzena Żygis, Susanne Fuchs

Leibniz-Centre General Linguistics (Leibniz-ZAS)
zygis@leibniz-zas.de, fuchs@leibniz-zas.de

ABSTRACT

Speakers flexibly adapt their speech production to situational demands. This study investigates the extent to which speakers adjust their articulation in terms of lip aperture depending on (i) the speech mode (normal speech vs. whispered speech where f0 is absent), (ii) the visibility of the interlocutor (visible vs. invisible), and (iii) the pragmatic function of the message (question vs. statement). To this end, maximal lip aperture in German vowels was scrutinized by means of a motion capture experiment.

Based on ten speakers, our results reveal that lip aperture is larger (i) in whispered than in normal speech, (ii) when speakers do not see each other, and (iii) also when questions rather than statements are being produced.

All the results suggest trade-off relations where the lack of both fundamental frequency and visibility are compensated for by larger lip aperture.

Keywords: lip aperture, whispered speech, speaker visibility, prosody, German

1. INTRODUCTION

Trading relations, in which one cue compensates for the absence or reduced occurrence of another, have been widely discussed in audiovisual perception (e.g. [5], [13], [10]), in speech production with respect to idiosyncratic properties and motor equivalence [14], and also in terms of speech and hand gestures ([8]). According to the *trade-off* hypothesis, if situational constraints demand it, different modalities can be used to compensate for another. For example, interacting in a noisy environment may lead to the use of enlarged articulatory and nonverbal gestures to compensate for reduced auditory information by enhancing visual information ([4], [17]).

Although several previous studies have investigated trade-off relations, their object of investigation has been limited to voiced speech [5, 6]. It remains unclear, however, what happens to articulatory gestures when the acoustic speech signal becomes voiceless (e.g. due to whispering) and therefore harder to understand. More specifically, we address the question of whether lip aperture also enters a trade-off relation with speech signal when the latter is deprived of one of its underlying cues – namely, fundamental frequency [12]. Are articulatory or facial gestures then more pronounced, contributing more intensively to the understanding of speech? In this paper we will concentrate on articulatory gestures only, but further work is in progress to integrate facial gestures such as eyebrow motion as well.

Questions regarding whispered speech and trade-off relations have thus far scarcely been addressed in the literature, with the exception of [6], which revealed that perception of prosodic focus in French whispered speech is difficult to discern when based on acoustic signal only. In addition, reaction time measurements showed that when acoustic cues are not sufficient, as is the case with whispered speech, adding vision (i.e. oro-facial expressions) decidedly enhances perception of prosodic focus, leading to much faster reaction times.

Our study differs from [6] in that it does not investigate perception but production of whispered speech. In particular, it sets out to examine lip aperture in whispered as opposed to normal speech mode. We hypothesize that the lip aperture is larger in whispered than in normal speech, thereby compensating for the degraded speech signal.

Previous research has also investigated whether the relation between speech and gestures is in function of mutual visibility. Again, several studies have concentrated on hand gestures (see [2] for a summary) and shown that the gesture rate decreases when interlocutors are not mutually visible.

Our study extends the repertoire of gestures by examining lip aperture, which is an obligatory articulatory gesture and thus differs in nature from e.g. hand gestures. Our aim is also to help to understand whether lip aperture varies depending on situational context. In this regard, our goal is to investigate whether participants speak with a larger lip opening when not visible to each other.

Furthermore, as shown in our previous study [17], lip aperture is larger in questions than in statements. The present study takes this pragmatic factor into account by creating a more ecologically valid setting, in which the sentences are produced during the interaction with the interlocutor (and not read from a computer screen as in [17]). In addition, the production of the sentences is investigated in the

visible/invisible mode.

Thus, the present study will fill a research gap by investigating lip aperture focusing on three factors:

(a) the speech mode (whispered vs. normal speech)
(b) the visibility of the interlocutor (visible vs. invisible mode)
(c) the pragmatic function of a message related to prosody (polar questions with rising F0 vs. statements with non-rising F0).

The investigation of these factors will help us to scrutinize the nature of lip opening as an articulatory gesture.

2. EXPERIMENT

2.1. Informants and experimental design

To meet our research goals, we conducted a motion capture experiment with ten native speakers of German (six female speakers, mean age 29.2 (5.71 s.d.)).

To measure lip aperture (and eyebrow movements), seven markers were placed on the face in the following way: four markers around the lips, i.e. (i) below the lower lip so that the marker was not hidden by it, (ii) above the upper lip, (iii) at the left lip corner, and (iv) at the right lip corner. One marker was placed above the nose in the central position between the eyebrows. This marker, as well as three additional markers fixed on glasses frames, served as reference points. (Two other markers were put slightly above the left and right eyebrow.) Figure 1 illustrates the positions of the markers.

Figure 1: Positions of facial markers (glasses were originally transparent; they are painted black here for anonymization).

The recordings were obtained by means of a motion capture system (OptiTrack, Motive Version 1.9.0) with 12 cameras (Prime 13) in a sound-proof lab. Motion data was recorded with a sampling frequency of 200 Hz. The parallel acoustic recordings were conducted using a Sennheiser ME62 microphone (20 cm distance from lips) at a sampling rate of 44100 Hz.

The task of the informant was to respond with (a) a question mirroring the statement uttered by the confederate, or (b) a statement mirroring the question asked by the confederate. The participant changed neither the content of the sentence nor its word order. Instead they altered their intonation while producing the same sentence; see examples in (1a, b).

(1a) Question condition
Confederate: *Er mag diese Piste.*
"He likes this slope."
Informant: *Er mag diese Piste?*
"He likes this slope?"

(1b) Statement condition
Confederate: *Er mag diese Piste?*
"He likes this slope?"
Informant: *Er mag diese Piste.*
"He likes this slope."

There were 40 sentences, i.e. 20 pairs of statements and questions. In their final positions they included strictly controlled words which were always bisyllabic with stress falling on the first syllable – such as, for instance, *Mandel* "almond", *Männer* "men", *Pasta* "pasta", *Pelze* "furs", *Matte* "mat", *Bitte* "request", *Masse* "mass", *Bälle* "balls", *Bände* "volumes". All words started with a bilabial stop /p/, /b/, /m/ followed by /a/, /ɛ/ or /ɪ/ and the syllables always had a CVC structure. The advantage of bilabial stops is that they involve lip closure in their articulatory realization, prior to a lip aperture for the following vowel. All vowels were unrounded, but differed in their height: from the greatest aperture in the case of /a/ to the smallest aperture in the case of /ɪ/.

As far as intonation pattern in German is concerned, the nuclear accent falls on the sentence's final content word, i.e. the last stressed syllable in an Intonational Phrase (corresponding to a sentence in our case). The boundary tone is high (H%) in polar questions and low in statements (L%) [7]. In both sentence types examined, the final part of the sentence was a content word carrying an accent, so the intonation contour included a pitch accent on the word and a boundary tone reflecting the question vs. statement distinction (at least in voiced speech).

The experiment consisted of the following four stimuli blocks, as presented in (2). Each block contained questions and statements. The sentences were randomized and three repetitions of the randomized lists were conducted. The order of block presentations was also randomized for each speaker.

(2) Blocks:
a) normal speech, informants see each other
b) whispered speech, informants see each other
c) normal speech, informants do not see each other
d) whispered speech, informants do not see each other

In order to elicit the data in the invisible mode, the confederate and the informant were separated by an artificial wall as shown in Figure 2.

Figure 2: Experimental setting for the invisible mode (with an artificial wall between the speakers)

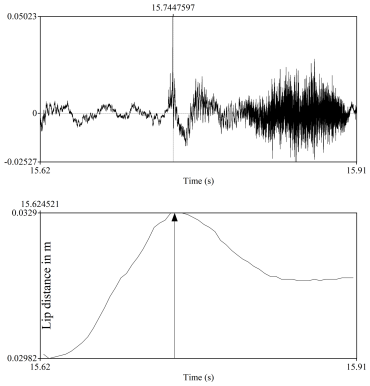
In total, we analyzed 1554 items with respect to the lip aperture (40 sentences x 3 repetitions x 4 speech modi x 2 visibility conditions x 10 speakers); 48 items were not examined for various reasons.

2.2. Annotation and analyses

For the purposes of the present study, we acoustically labeled the target word from the beginning of the closure to the end of the word using Praat 6.0.40 ([3]). Five timepoints were determined from the spectrogram: (i) the onset of the stop phase of the word-initial stop, (ii) the onset of the stop burst, (iii) the onset of the vowel, (iv) the offset of the vowel, and (v) the offset of the word.

These temporal landmarks were used to manually determine the minimum and maximum of 3D lip distance from the bilabial to the vowel using MATLAB [11]; see Figure 3.

Figure 3: Measurements of lip distance in the production of the stop and following vowel



Distances between the different markers were calculated as follows:

(2) (dist=sqrt((x_marker1- x_marker2)^2 +(y_marker1-y_marker2)^2+(z_marker1-z_marker2)^2)

2.3. Statistics

Linear mixed effect models were employed for assessing the influence of SPEECH MODE [normal, whispered], VISIBILITY MODE [visible, invisible], SENTENCE TYPE [question, statement] and VOWEL TYPE [a, ɛ, ɪ] on LIP APERTURE, as well as their interaction. The Type I error was minimized by using the maximized structure (see [1]). However, due to high correlations found between random-effect terms, most random structure was removed and the final model included speaker-specific intercept and slope for the sentence type and word intercept and slope for speech mode (no high correlations between fixed effects were observed.). The maximized models were tested against less complex models by means of likelihood ratio tests, and the best fit model was selected as the final one. The p-values were estimated with the Satterthwaite approximation with the help of lmerTest ([9]).

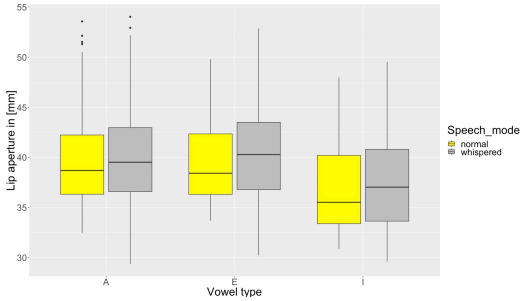
All statistical analyses were conducted in the R Studio software (version 1.1.453 [15]).

3. RESULTS

Our results reveal that all vowels are produced with a larger lip opening in whispered than in normal speech (t=3.18, p<.001). Note that we present our results by considering inherent lip opening differences for vowels: from /a/ with the largest lip opening to /ɪ/ with the smallest lip opening.

The lip aperture differs for individual vowels: the vowels [ɛ] and [a] are produced with a significantly larger lip opening than [ɪ]. However, there is no difference in lip aperture between [ɛ] and [a] for both whispered and normal speech (whispered speech: [ɛ] vs. [i] t=7.42, t=.001, [a] vs. [ɪ] t=7.63, p<.001; normal speech: [ɛ] vs. [ɪ] t=8.66, p<.001, [a] vs. [ɪ] t=11.75, p<.001). The interaction Vowel type*Speech mode is also significant (t=2.57, p<.05), indicating differences in lip opening between normal and whispered speech for individual vowels, as presented in Figure 4.

Figure 4: Lip aperture in different vowel as function of speech mode



As far as the visibility mode is concerned, the lip opening in all vowels is larger when the speakers do not see each other (t=-2.05, p<.05). If we compare individual vowels, a similar grouping to before emerges: vowels [ɛ] and [a] are produced with a significantly larger lip opening than [ɪ] ([ɛ] vs. [i]: t=7.19, t=.001; [a] vs. [ɪ]: t=6.75, p<.001). The interaction of Visibility*Vowel is not significant, which suggests that differences between the visible vs. invisible mode are similar for all vowels. This is illustrated in Figure 5.

Figure 5: Lip aperture in different vowels as function of visibility mode

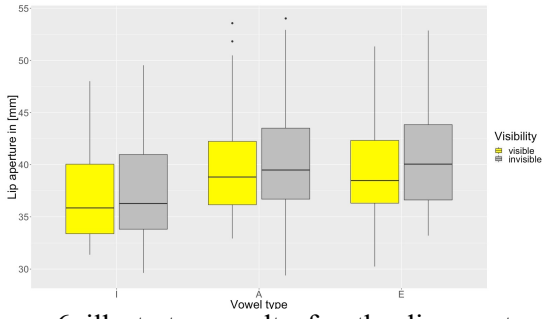
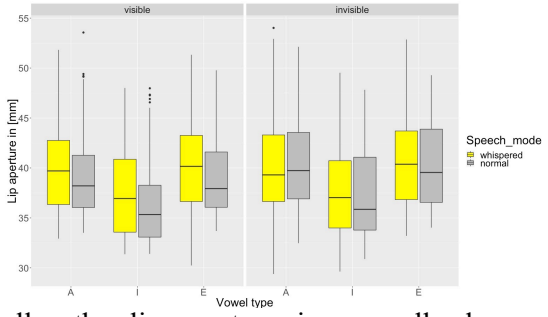


Figure 6 illustrates results for the lip aperture in individual vowels split according to visibility and speech mode. Whereas the aperture is higher for all vowels in visible mode, no difference is found for the vowel [a] in the invisible speech mode. The interaction Speech mode*Visibility is significant ($t=2.98$, $p<.01$).

Figure 6: Lip aperture in different vowels as function of visibility mode



Finally, the lip aperture is generally larger for questions than for statements ($t= 3.08$, $p<.01$). More specifically, the lip aperture is larger in questions for vowels [a] and [ɛ] than for the vowel [ɪ] ([a] vs. [ɪ]: $t=11.7$, $p<.001$); [ɛ] vs. [ɪ]: $t=8.66$, $p<.001$). The difference between [a] and [ɛ] is not significant. In addition, the interaction Sentence*Vowel type is also significant ($t=2.37$, $p<.05$), which points to differences in lip aperture between questions and statements for the three vowels, as illustrated in Figure 7.

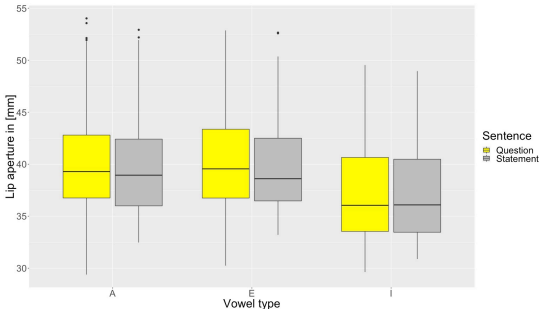
4. DISCUSSION AND CONCLUSIONS

Our results point to three main conclusions. First, lip aperture is larger in whispered as opposed to normal speech. Second, it is also larger when speakers do not see each other. Finally, vowels are articulated with a larger lip aperture when questions as opposed to statements are pronounced.

As far as individual vowels are concerned, [a] and [ɛ] pattern together by showing a greater lip aperture than [ɪ].

Regarding trade-off relations, our results suggest compensation effects between degraded acoustic signals and articulatory gestures on various levels:

Figure 7: Lip aperture in different vowels as function of sentence type



1) the lack of fundamental frequency is compensated for by a larger lip aperture, which may enhance visual cues for the interlocutor (but also affect the acoustic signal)
2) the lack of the visibility of the interlocutor is also compensated for by a larger lip aperture.

The results support the *trade-off* hypothesis in a sense that differences in prosody (question vs. statement) are executed in whispered speech also by a larger lip opening. However, the underlying mechanisms triggering the dependence of rising intonation and larger lip opening require further investigation.

Furthermore, the study provides an answer to the question of whether the speaker uses lip opening to enhance visual perception for the interlocutor or whether the lip opening helps the speaker to transmit the message, with the greater effort put in enhancing the acoustic properties of higher frequency (e.g. formants). Given that lip opening is larger in the invisible condition where the speaker does not see her interlocutor, the second explanation may be favored. Hence, lip opening is an obligatory articulatory gesture which might be enhanced depending on the situational context.

In summary, a complex picture of a compensatory multimodal interaction emerges when speech mode and speaker visibility are taken into account, with the common denominator of larger lip aperture compensating for the lack of acoustic (f0) or visual cues given certain situational demands. In light of this, the result can also be explained in terms of H&H theory [10], where trade-off relations are assumed between listener comprehension and speaker production depending on the situational context.

Acknowledgements

This research has been supported by the Bundesministerium für Bildung und Forschung (BMBF, Germany) Grant Nr. 01UG1411 and Leibniz Society. We also thank Jörg Dreyer for technical support.

5. REFERENCES

[1] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 255-278.

[2] Bavelas, J., Gerwing, J., Sutton, C., Prevost D. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language* 58, 495-520.

[3] 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/

[4] De Ruiter, J. P., Bangerter, A., Dings, P. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science* 4(2), 232–248.

[5] Diehl, R. L. 2011. On the robustness of speech perception. *Proceedings of the XVII International Congress of Phonetic Sciences*. University of Hong Kong, 1-8.

[6] Dohen, M., Lœvenbruck, H. 2009. Interaction of audition and vision for the perception of prosodic contrastive focus. *Language & Speech* 52, 177-206.

[7] Grice, M., Baumann, S., Benzmüller, R. 2005. German Intonation in Autosegmental-Metrical Phonology. In S.-A. Jun (ed.). *Prosodic Typology: The Phonology of Intonation and Phrasing* Oxford, UK: Oxford University Press. 55-83.

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

[9] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2013. lmertest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0–25.

[10] Lindblom, B. 1990. *Explaining phonetic variation: A sketch of the H&H theory*. In: Hardcastle, W.J., Marchal, A. (eds.) *Speech production and speech modelling*. Springer, Dordrecht. 403-439.

[11] MathWorks 2017. MATLAB. Natick, Massachusetts, USA.

[12] Miller, G. A., Nicely P. 1955. An Analysis of Perceptual Confusions among some English Consonants. *Journal of the Acoustical Society of America* 27(2), 338-352.

[13] Parker, E. M., Diehl, R. L., Kluender, K. R. 1986. Trading relations in speech and nonspeech. *Perception & Psychophysics* 39, 129–142.

[14] Perkell, J. S., Matthies, M. L., Svirsky, M. A., Jordan, M. I. 1993. Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot ‘‘motor equivalence’’ study. *The Journal of the Acoustical Society of America,* 93(5), 2948-2961.

[15] RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. Version 1.1.453.

[16] Van der Sluis, I., Krahmer E. 2007. Generating multimodal references. *Discourse Processes* 44, 145–174.

[17] Żygis, M., Fuchs, S., Stoltmann, K. 2017. Orofacial expressions in German questions and statements in voiced and whispered speech. *Journal of Multimodal Communication Studies* 4, 87-92.