# DISCREPANT COMBINATIONS OF ACOUSTIC AND VISUAL SPEECH SIGNALS AND ITS RELATION TO ACOUSTIC DISTANCE IN VOWELS

Risa Matsubara[1], Terumichi Ariga[1], Takeshi Kishiyama[1], Chuyu Huang[2]

The University of Tokyo[1], Nagoya Gakuin University[2]
matsubara-risa761@g.ecc.u-tokyo.ac.jp, ariga@phiz.c.u-tokyo.ac.jp, kishiyama.t@gmail.com, huang@ngu.ac.jp

## ABSTRACT

This study aims to investigate two research questions: (1). whether and how the acoustic characteristics affect the audiovisual perception of Estonian back-mid unrounded vowel [ɤ], which posits between [ø] and [o] in terms of acoustic characteristics, and (2). if the listeners visually judge the front or back of the vowel. The accuracy rate of the audiovisual ABX task was significantly lower for the combination of the audio information of [ɤ] with the visual information of [ø] or [o] but higher when the auditorily presented vowel and the visually presented vowel were opposing one another in terms of the backness (e.g., audio: [o] and visual: [ø]). The results suggest that the perceptual illusion of vowels is more likely to occur when the acoustic gap between the auditorily presented vowel and the visually presented vowel is small, even if the listeners have acoustically similar vowels in their native language.

**Keywords:** audiovisual speech perception, McGurk effect, Estonian, vowel perception, lipreading.

## 1. INTRODUCTION

Humans perceive speech sounds not only through audio information but also through visual information. This phenomenon is known as audiovisual integration, and many studies have examined the role of visual information in perceiving speech sounds [1–4]. Listeners utilize visual information in speech perception because it provides information about articulatory motions that can act as a cue to identify speech segments [1]. When audio speech perception is difficult, listeners attend more to talkers' mouth [2, 3]. Visual information, therefore, helps listeners to identify speech perception in adverse conditions.

However, audiovisual integration in speech perception can cause listeners to perceive illusional sounds that are not actually presented in an audio mode. A famous phenomenon is the McGurk effect, in which perception is distorted by the gap between visual and audio information [4]. For example, when the audio of the sound [ba] is combined with a video where a person is pronouncing [ga], the resulting information is perceived as [da].

The same phenomenon is also true when it comes to vowel perception. For example, Traummüller & Öhrström [5] reported that the combination of the audio information of an unrounded front mid vowel [e:] and the visual information of a rounded front high vowel [y:] was perceived as [ø:], a rounded front-mid vowel, by native speakers of Swedish (see Engstrand [6] for the vowel inventory of Swedish). They concluded that audio information (e.g., [e]: [–round][mid]) provides information about height while visual information (e.g., [y]: [+round][high]) offers information about the roundedness of the vowel. Accordingly, the results imply that both audio and visual information affect the perception of vowels, while roundedness is likely to be visually perceived.

However, it is possible that the listeners depend solely on the audio information when judging the roundedness of the vowel. Lisker and Rossi [7] instructed French listeners to judge the roundedness of the vowel of a video in which the audio information was discrepant with the visual information in terms of the roundedness. The results suggested that the audio information of rounded vowels (e.g., [ø]) was more likely to be perceived as [+round] than the visual information of rounded vowels. This was interpreted to mean that the listeners can shift to one input mode despite the discrepancy between the audio information and visual information. Additionally, a higher misperception rate of [ɨ] or [ɯ] as [+round] when presented in combination with the image of rounded vowels was shown, which implies that central or back unrounded vowels are close to rounded vowels in terms of their acoustic quality. In fact, [ɨ] has a closer acoustic value to that of [y] [7]. However, what should be noted here is that French listeners do not have an inventory of unrounded central or back vowels although they have front and back rounded vowels such as [y], [u], or [ø] [8]. Therefore, the reason that they perceived

the unrounded vowels [ɨ] or [ɯ] as [+round] may be the result of the assimilation of the unfamiliar vowels into their familiar category of rounded vowels based on acoustic and perceptual information. This study aims to test if such misperception of unrounded vowels as rounded ones in audiovisual perception still occurs when the listeners have both rounded front and back vowels and an unrounded back vowel at the same height in their vowel inventory.

Estonian has an unrounded back vowel [ɤ] as well as a rounded front vowel [ø] and a rounded back vowel [o], both of which are at the same height as the [ɤ]. The vowel [ø] has a higher second formant (F2) value, which represents frontness, while the vowel [o] has a lower F2 value, reflecting backness. Both [ø] and [o] have a close first formant (F1) value, indicating their similarity in height.

The vowel [ɤ] acoustically and perceptually sits between the rounded front [ø] and the rounded back [o] (Figure 1). This means that [ɤ] can be perceived as [+round] due to its acoustic characteristics when combined with the visual information that provides [+round]. However, if having the two acoustically similar but articulatory different vowels in their native language enables the listeners to focus on the acoustic information regardless of the acoustic similarity, [ɤ] will be correctly judged even when combined with [+round] vowel. Additionally, whether the listeners can visually perceive the backness of the vowel is unclear. In the case of Estonian, [ø] and [o] are the same in terms of the roundedness, but they differ in terms of backness (Table 1). If the backness is visually perceived as well as the roundedness, the listeners will perceive audio [o] as [ø] when combined with the visual information of [ø]. Alternatively, if the listeners depend more on the audio information when judging the backness due to the larger acoustic distance between the auditorily presented vowel and the visually presented vowel, they will correctly perceive the audio [o] even when combined with [ø].

**Table 1:** The features of the Estonian vowels [ø], [ɤ], and [o].

| | round | back |
|---|---|---|
| ø | + | − |
| ɤ | − | + |
| o | + | + |

Thus, this study investigated whether the visual information of [+round] affects the perception of the acoustically vague vowel and whether the listeners visually judge the backness. To test the research questions, we conducted an audiovisual

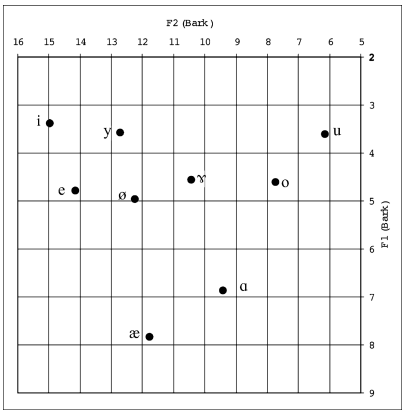


**Figure 1:** The acoustic and perceptual space of Estonian vowels (cited from Asu and Teras [9]).

ABX experiment, where the listeners needed to judge which third stimulus (X) was identical to either of the first two, A or B. If the listeners make more errors in their judgement, it means that they have difficulty determining the phonetic value of the stimulus. There may be an increase in their reaction times (RTs), which reflects the difficulty of their judgement as well.

## 2. METHOD

### 2.1. Participants

Thirty-two native speakers of Estonians were recruited using the website Prolific [10]. They indicated that they were raised only using the Estonian language.

### 2.2. Materials

Four Estonian vowels [e, o, ø, ɤ] were pronounced by three native speakers of Estonian (two males and one female; mean age = 39 years). They were recorded via Zoom [11] with a sampling rate of 32.0 kHz. They were instructed to pronounce the vowels in the same natural pitch pattern. The audio and visual information was separated and then combined by using iMovie to produce the videos in which the audio information corresponded with the visual information or did not. The acoustic information of the vowels they produced is shown in Table 2.

To avoid the effect of acoustic differences other than vocalic differences, the absolute value of amplitude was set to 0.7 after modification of the intensity to 70 dB in Praat [12].

There were two items, as in Table 3 and 4. In Video A, the audio vowel [ɤ] was combined with either the visual information of [o] or [ø], the acoustically close vowels, to observe if the

**Table 2:** Acoustic information of the vowels (Hz). S1 = Speaker 1, S2 = Speaker 2, and S3 =Speaker 3. M is an abbreviation of "male" and F is an abbreviation of "female". F1 = first formant, F2 = second formant, F3 = third formant.

| | Formant | S1 (M) | S2 (M) | S3 (F) |
|---|---|---|---|---|
| e | F1 | 428.1 | 402.9 | 373.8 |
| | F2 | 2068.4 | 2299.6 | 2632.5 |
| | F3 | 2390.2 | 2860.6 | 2968.3 |
| o | F1 | 491.6 | 416.32 | 418.9 |
| | F2 | 636.3 | 628.9 | 608.4 |
| | F3 | 2244.1 | 2894.9 | 2807.7 |
| ø | F1 | 443.2 | 419.1 | 432.3 |
| | F2 | 1497.3 | 1475.8 | 2192 |
| | F3 | 2150 | 2383.2 | 2756.6 |
| ɤ | F1 | 429.59 | 398.1 | 434.2 |
| | F2 | 1307.6 | 1067.9 | 1083.6 |
| | F3 | 2261.3 | 2496.9 | 2964.6 |

perception is influenced by the visual [+round] information. Video B was a control in which there was no incongruency between audio and visual information.

**Table 3:** An example of items (audio [ø]).

| | Video A | Video B |
|---|---|---|
| Experimental | audio [ɤ] + visual [ø] | audio [ø] + visual [ø] |
| Backness | audio [o] + visual [ø] | audio [ø] + visual [ø] |
| Control | audio [ɤ] + visual [ɤ] | audio [ø] + visual [ø] |

**Table 4:** An example of items (audio [ɤ]).

| | Video A | Video B |
|---|---|---|
| Experimental | audio [ɤ] + visual [o] | audio [o] + visual [o] |
| Backness | audio [ø] + visual [o] | audio [o] + visual [o] |
| Control | audio [ɤ] + visual [ɤ] | audio [o] + visual [o] |

The order of the stimuli and the speakers was counterbalanced to avoid the effect of the difference in the presentation order of the two videos. The number of the experimental stimuli, therefore, totaled 48 (3: conditions x 2: items x 2: speaker identity x 2: order x 2: X identity). Speaker 3 was assigned to the sound X in all stimuli while the speaker of Video A or B were either Speaker 1 or Speaker 2.

### 2.3. Procedure

The experiment was created on PCIbex [13]. The participants remotely took part in the experiment by clicking the link. They listened to the materials with headphones or earphones at a comfortable volume.

The task was an audiovisual ABX task in which stimuli X were presented only auditorily. The two videos, A and B, were played in a row and only the sound that was identical to either A or B was played at the end. The participants were required to judge which sound was identical to the last sound X by pressing the A key or the B key. The maximum time to answer was 4000 ms. The 48 stimuli were presented randomly along with another 48 filler stimuli. The total time needed was 30 minutes, and all of the participants were paid for their participation.

### 2.4. Analysis

Before proceeding to the statistical analysis, data on two participants were excluded from the data because of their extremely lower accuracy rate when compared to other participants (55% and 65% respectively). Judging by the histogram, the tokens whose reaction times (RTs) were no more than 150 ms or more than 3000 ms were omitted from the data. After the data trimmed, the data that exceeded $\pm 2.5SD$ when based on the log-transformed RTs were excluded.

The statistical analysis involved the generalized linear mixed-effects models (GLMM) and linear mixed-effects (LME) models. The GLMM was used for the analysis of the accuracy rate and the LME model was used for the analysis of the RTs.

In the GLMM analysis, the participants' answers (correct or wrong) were set as a response variable, the condition (Experimental, Backness, or Control) was set as a fixed variable, and the individual differences in the items and participants were set as random effects. In the LME analysis, RTs were set as a response variable, the condition (Experimental, Backness, or Control) was set as a fixed variable, and the individual differences in the items and participants were set as random effects. The best-fit model was selected by using backward selection [14].

## 3. RESULTS

The mean accuracy rates per condition are shown in Table 5.

**Table 5:** The mean accuracy rate (%).

| | accuracy (%) |
|---|---|
| Control | 90.4 |
| Experimental | 85.3 |
| Backness | 93.6 |

The GLMM analysis found a significant difference between the Experimental and the Control conditions and a marginally significant

difference between the Backness and Control conditions (Table 6).

**Table 6:** The statistical results of the accuracy rate.

| | *β* | *SE* | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.5205 | 0.3838 | 6.567 | < 5.14e–11 | *** |
| Experimental | –0.5496 | 0.2134 | –2.576 | 0.010 | ** |
| Backness | 0.4525 | 0.2506 | 1.806 | 0.071 | . |

However, there were no significant differences in the RTs between the conditions (Table 7).

**Table 7:** The statistical results of the RTs.

| | *β* | *SE* | *df* | *t* | *p* | |
|---|---|---|---|---|---|---|
| (Intercept) | 621.99 | 35.04 | 22.90 | 17.750 | < 7.02e–15 | *** |
| Experimental | 24.66 | 29.13 | 1307.42 | 0.846 | 0.397 | |
| Backness | −29.41 | 28.83 | 1307.12 | −1.020 | 0.308 | |

## 4. DISCUSSION

The present study tested these two hypotheses. The first hypothesis was that the visual [+round] information changes the perception of a [−round] vowel to [+round] when they have similar acoustic characteristics even if both are included in the vowel inventory of the listeners' native language. The significantly lower accuracy rate in the Experimental condition can be interpreted to mean that the Estonian listeners perceived the [ɤ] as [+round] and supports the hypothesis that the audio perception of the vowel is distorted by the visual information when the acoustic information of vowels is close between the audio and visual materials. The observation that the unrounded vowel was perceived as a rounded vowel affected by the visual roundedness is in alignment with the results of Traummüller & Öhrström's [5] study. Moreover, the results show that having both an unrounded back vowel and rounded front and back vowels in the listeners' native language does not necessarily guarantee resistance to the effect of discrepant visual information of the rounded vowel on the perception of the auditorily presented unrounded back vowel, which is acoustically close to the rounded vowel.

The other hypothesis was that the listeners visually judge the backness regardless of the acoustic distance between the audio and visual information. However, the accuracy rate was not significantly different between the Backness and Control conditions, thus providing no support for this hypothesis. In fact, the accuracy rate was higher in the Backness condition than in the Control condition, although the difference was marginally significant. This implies that the effect of the visual information is eliminated and the listeners can correctly auditorily perceive the vowel when the acoustic difference is large between the audio and visual vowels and between the first and second stimuli. Consequently, the results weakly reject the hypothesis that the backness of vowels is visually judged.

Overall, this current study implied that the audiovisual illusion of the vowels was not language-specific. The audiovisual illusion occurred not only in Swedish or French, which has rounded vowels at both back and front, but also in Estonian, which has an unrounded back vowel in addition to the rounded front and back vowels. Also, the results add to those of Navarra and Soto-Faraco [15], who reported that visual information facilitates the perception of nonnative vowels, by demonstrating that the visual information of a discrepant vowel can negatively affect the perception of native vowels.

## 5. REFERENCES


[1] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychological Review*, vol. 74, no. 6, pp. 431–461, 1967.

[2] E. Vatikiotis-Bateson, I. M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception and Psychophysics*, vol. 60, no. 6, pp. 926–940, 1998.

[3] J. Birulés, L. Bosch, F. Pons, and D. J. Lewkowicz, "Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech," *Language, Cognition and Neuroscience*, vol. 35, no. 10, pp. 1314–1325, 2020.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[5] H. Traummüller and N. Öhrström, "Audiovisual perception of openness and lip rounding in front vowels," *Journal of Phonetics*, vol. 35, no. 2, pp. 244–258, 2007.

[6] O. Engstrand, "Swedish," *Journal of the International Phonetic Association*, vol. 20, no. 1, pp. 42–44, 1990.

[7] L. Lisker and M. Rossi, "Auditory and visual cueing of the [±rounded] feature of vowels," *Language and Speech*, vol. 35, no. 4, pp. 391–417, 1992.

[8] B. Collins and I. M. Mees, *Practical phonetics and phonology: A resource book for students*. Routledge, 2013.

[9] E. Asu and P. Teras, "Estonian," *Journal of the International Phonetic Association*, vol. 39, no. 3, pp. 367–372, 2009.

[10] "Prolific," https://www.prolific.co/, visited 6-Jan-23.

[11] "Zoom: One platform to connect," https://zoom.us/, visited 6-Jan-23.

[12] “Praat: Doing phonetics by computer,” https://www.praat.org/, visited 6-Jan-23.
[13] “PCIbex,” https://www.pcibex.net/, visited 6-Jan-23.
[14] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, “Parsimonious mixed models,” *arXiv preprint arXiv:1506.04967*., 2015.
[15] J. Navarra and S. Soto-Faraco, “Hearing lips in a second language: visual articulatory information enables the perception of second language sounds,” *Psychological Research*, vol. 71, pp. 4–12, 2007.