JSLHR

Article

Audiovisual Perception of Congruent and Incongruent Dutch Front Vowels

Bea Valkenier,[a,b] Jurriaan Y. Duyne,[b,c] Tjeerd C. Andringa,[a,b] and Deniz Başkent[a,b]

Purpose: Auditory perception of vowels in background noise is enhanced when combined with visually perceived speech features. The objective of this study was to investigate whether the influence of visual cues on vowel perception extends to incongruent vowels, in a manner similar to the McGurk effect observed with consonants.
Method: Identification of Dutch front vowels /i, y, e, Y/ that share all features other than height and lip-rounding was measured for congruent and incongruent audiovisual conditions. The audio channel was systematically degraded by adding noise, increasing the reliance on visual cues.
Results: The height feature was more robustly carried over through the auditory channel and the lip-rounding feature through the visual channel. Hence, congruent audiovisual presentation enhanced identification, while incongruent presentation led to perceptual fusions and thus decreased identification.
Conclusions: Visual cues influence the identification of congruent as well as incongruent audiovisual vowels. Incongruent visual information results in perceptual fusions, demonstrating that the McGurk effect can be instigated by long phonemes such as vowels. This result extends to the incongruent presentation of the visually less reliably perceived height. The findings stress the importance of audiovisual congruency in communication devices, such as cochlear implants and videoconferencing tools, where the auditory signal could be degraded.

Key Words: audiovisual speech perception, vowels, McGurk effect

Perception of spoken language is not an auditory phenomenon only; it is also heavily influenced by visually perceived pronunciation information. The influence of visual cues on speech perception has been shown for a variety of speech tokens such as consonants (see, e.g., Massaro, 1987; for an overview, see Massaro, 1989; Massaro & Cohen, 1990) and vowels (Robert-Ribes, Schwartz, Lallouache & Escudier, 1998; Traunmüller & Öhrström, 2007) and for conditions such as hearing impairment (Başkent & Bazo, 2011; Grant, Walden, & Seitz, 1998; Miller & D'Esposito, 2005). This interaction is so strong that when the auditory and visual components are incongruent, they may fuse into a single percept that is different from both the original auditory stimuli and the original visual stimuli—also known as the *McGurk effect* (McGurk & MacDonald, 1976). For spoken man–machine interaction devices and video applications, such knowledge of audiovisual integration is crucial. For example, the precision with which the auditory and visual information are aligned in videoconferencing tools follows directly from research on audiovisual integration of temporally mismatching stimuli (McGrath & Summerfield, 1985; Miller & D'Esposito, 2005). Also, appropriate audiovisual alignment is especially important for users of rehabilitative communication devices such as cochlear implants and hearing aids. Because the auditory signals are less well transmitted, listeners with hearing impairment rely heavily on the visual cues (Başkent & Bazo, 2011; Champoux, Lepore, Gagneú, & Théoret, 2009; Rouger, Fraysse, Deguine & Barone, 2008). When auditory information is correctly aligned with visual information, listeners—especially those with hearing impairment—profit significantly from the visual information for understanding speech (Başkent & Bazo, 2011). However, when audiovisual information is not correctly aligned, disruptive interactions may be observed in addition to the loss of positive interaction. Disruptive interactions of audiovisual information have been shown with the McGurk effect for consonants but are not as extensively investigated for the case of vowels. However, it was recently shown that the contribution of vowels to the auditory intelligibility

[a]University of Groningen, the Netherlands
[b]University Medical Center Groningen
[c]University of Cambridge, United Kingdom
Correspondence to Bea Valkenier: B.Valkenier@ai.rug.nl
Editor: Sid Bacon
Associate Editor: Charissa Lansing
Received August 15, 2011
Revision received February 3, 2012
Accepted April 26, 2012
DOI: 10.1044/1092-4388(2012/11-0227)

of speech is significant and could, in some listening situations, be even higher than the contribution of consonants (Cole, Yan, Mak, Fanty, & Bailey, 1996; Kewley-Port, Burkle, & Lee, 2007). Kewley-Port et al. (2007) argued that listeners with hearing impairment are even more dependent on the correct perception of vowels because, in most cases of hearing impairment, high frequencies (which are associated with consonants) are lost more readily than are low frequencies (which are associated with vowels). Thus, correct alignment is shown to be important for audiovisual interaction devices, and although vowels are shown to be important for speech intelligibility, research has focused on audiovisual incongruence with consonants. As vowels are of higher intensity and have longer duration than consonants, the effect of visually incongruent information—for example, as in cochlear implant or hearing aid users—might be different for vowels than for consonants. In the present study, therefore, we investigated the perceptual processes that play a role in the audiovisual perception of vowels—more specifically, the Dutch high- and mid-high–front vowels ([i, y, e, Y], as in the Dutch words *biet, fuut, beet,* and *hut,* respectively)—with congruent and incongruent audiovisual features.

On the basis of acoustic information, the first formant (F1) and second formant (F2) of a particular vowel are most crucial for its recognition (for an overview, see Rosner & Pickering, 1994). Regarding the vowels of interest of the present study, the F1 is generally associated with the height feature, and the F2 is generally associated with the backness feature (Ladefoged, 1982; Rosner & Pickering, 1994). Furthermore, the literature suggests that F2 is also related to the lip-rounding feature for some vowels (Lisker & Rossi, 1992; Valkenier & Gilbers, 2008). Masking one of the formants by noise leads to perceptual confusions. By establishing confusion matrices for different levels of white noise, Pickett (1957) showed that the shared features of the vowels explain the relatively structured confusions that were observed. In short, height—by virtue of the perception of F1—is the most robust acoustic feature, followed by backness (F2).

In addition to the acoustic cues, visual cues also influence the perception of high-front vowels. Robert-Ribes et al. (1998) quantified the facilitatory influence of visual cues on the French high- and mid-high–front vowels [i, y] and [e, ø] by using congruent audiovisual stimuli presented with white noise at different levels. In most cases, the visual and auditory cues are complementary (Massaro & Stork, 1995); for instance, lip rounding is a strong visual cue, whereas height is a strong auditory cue. Similarly, Miller and Nicely (1955) showed that most features of consonants that were easy to identify from a talker's face were hard to identify from hearing them, and vice versa. Summerfield (1987) labeled and described those findings as *complementarity in audiovisual processing.* Complementarity of the two modalities improves the perception of congruent audiovisual stimuli, especially when the auditory input is deteriorated (e.g., in background noise). However, if the audiovisual stimuli are incongruent, fusions may occur, such as in the McGurk effect (McGurk & MacDonald, 1976). In short, when a visual [ga] stimulus[1] was concurrently presented with an auditory [ba] stimulus, the resulting perception was that of /da/.[2] The McGurk effect is extensively investigated on different pairs of consonants. However, research has not yet established the limits and the magnitude of the fusion effect in the vowels that are acoustically more stable. One reason for this could be that such an investigation is relatively difficult to do in English, where the most visually distinctive feature—lip rounding—is not an independent distinctive feature of vowels. In other languages with an independent lip-rounding feature, however, an experiment can be conceived that uses vowels that share all perceptual features but rounding. In Swedish, for example, Traunmüller and Öhrström (2007) found a shift in the auditory response from the Swedish high unrounded front vowel /e/ to the high rounded front vowel /ø/, when an auditory [e] stimulus was shown concurrently with a visual [y] stimulus. However, this effect was not generalizable, as it was observed only with a subgroup of participants who were more prone toward using visual speech cues.

The aim of the present study was to establish the extent to which the acoustic and visual domains influence audiovisual vowel perception, both in quiet and in background noise. In addition to congruent audiovisual vowel perception, taking advantage of the lip-rounding feature of Dutch vowels, we investigated the perceptual fusion using incongruent audiovisual stimuli. If the visual and acoustic features are complementary, as argued by Robert-Ribes et al. (1998), the visually more salient (i.e., prominent) feature (i.e., lip rounding) leads to a stronger McGurk effect than the visually less salient one (i.e., height). For this purpose, we measured confusions (similar to the measurements done in Traunmüller & Öhrström, 2007, and Robert-Ribes et al., 1998) with the Dutch high– and mid-high–front vowels of [i, y, e, Y]. These vowels allowed vowel pairs that would differ only in height or lip rounding. Hence, in the incongruent stimuli, conditions of audio and video input that differed in height only and/or rounding only could be tested. Contrary to the findings of Traunmüller and Öhrström (2007)—who analyzed the data for the subset of participants who were more prone toward using visual cues—we included all participants without a pre-selection. We induced a visual bias—that is, an increased reliance on visual information in audiovisual perception—for all participants by systematically adding noise to the auditory channel. The advantage in doing so is that the results can now be generalized

[1]In this notation, square brackets are used for phones or specific utterances.
[2]In this notation, slashes indicate perceived vowel quality or perceived vowel class.

to not only the subgroup of perceivers who are more prone toward visual cues but also to the entire group of normal-hearing listeners and their audiovisual speech perception in suboptimal listening conditions.

Method

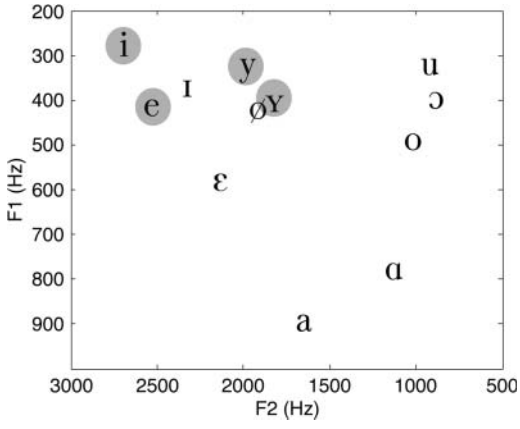
Subjects

Sixteen native speakers of Standard Dutch (11 men [M_{age} = 24.8 years, SD = 1.9]; 4 women [M_{age} = 23.8 years, SD = 0.5]) participated in the experiment. The data of one participant were not reliable because some data points were missing; therefore, we excluded all data of this participant from analysis. All participants reported normal hearing and normal-to-corrected vision. Participation was voluntary, with the possibility of withdrawing at any time during the study. Participants were fully informed about the study, and their written consent was obtained prior to data collection.

Stimuli

Selection of speech material and speech context. In order to give an impression of the Dutch vowel system, Figure 1 shows the vowel triangle of the Dutch vowels. The vowel triangle was created with the formants as determined with Praat (Boersma, 2001) from vowels produced in isolation by a 31-year-old female speaker of Standard Dutch. In the present study, we investigated the audiovisual perception of the Dutch high– and mid-high–front vowels [i, y, e, Y], which are represented as shaded circles in Figure 1. These vowels were selected because (a) lip-rounding and height features of these vowels cross in the acoustic as well as the visual domain and (b) there are no other confounding features. (For a more extensive analysis and justification of the selected vowels, see the Appendix.)

Figure 1. Vowel triangle of Dutch vowels produced in isolation by one female speaker. The shaded circles are the vowels that were used in this study.



The vowels [i, y, e, Y] were recorded in the context of [χ V χ], where [χ] represents a voiceless velar fricative (such as in the Dutch word *acht*). This choice was based on the argument by Traunmüller and Öhrström (2007) that velar consonants hardly affect the visibility of vowel features because the lips and the jaw do not need to be in a particular position. The voiced velar plosive [g], as was used in the Traunmüller and Öhrström study, does not exist in the Dutch language. Also, the use of the voiceless velar plosive [k] would lead to (semantically) meaningful Dutch words. As the context of the voiceless velar fricative χ produces phonologically allowed nonsense words for all Dutch vowels, this seemed to be the most appropriate context structure.

Recording and editing of speech material. The stimuli were recorded in a quiet room with bright natural daylight against a white background. The speaker was a 22-year-old female native speaker of the standard variety of Dutch. The stimuli were recorded with a Samsung HMX-H106-SP video recorder placed approximately 3 m from the speaker, who was standing against the white background with audio sampled at a rate of 48000 Hz. Recordings were made from the front of the speaker's face, including the entire face and neck and with the mouth at one third from the bottom of the display screen on the computer monitor. The total frame size on the computer monitor was 513 cm^2, and the size of the mouth was approximately 3 cm^2. The front portion of the tongue was visible for the high vowels (see Table 1).

For each vowel, two utterances were selected where the head movement was minimal, and the experimenters agreed on successful pronunciation of the target vowel. The duration of the video files of the selected stimuli were cut to equal duration of 1 s, with approximately 0.3 s neutral face at the start and end of the video. The long-term root-mean-square (RMS) levels of the audio-recordings were normalized with Praat. Steady low-pass filtered noise of varying intensities (SLN) was produced by low-pass filtering white noise (filter order = 1, resulting in a slope of –6 dB/oct in filter response). SLN was added to the stimuli (which were kept at constant intensity), resulting in stimuli with signal-to-noise ratios (SNRs; calculated on RMS levels) of 30 dB (almost quiet), 0 dB, –6 dB, –12 dB, and –18 dB. Audio presentation level of processed stimuli was calibrated to a comfortable level of approximately 70 dBA (varying with variations in stimulus intensity).

Those prepared recordings were used as control conditions and served as a starting point for the creation of the stimuli of the experimental conditions. In the experimental

Table 1. Articulatory features of the stimulus vowels.

Stimulus vowel	Lip-rounding	Height	Articulation example
/i/	Unrounded	High	
/e/	Unrounded	Mid-high	
/y/	Rounded	High	
/Y/	Rounded	Mid-high	

Note. Summary of the features of the Dutch front vowels used as stimuli with articulation example taken from the experimental stimuli (the section from around the mouth was cut from a still frame from the corresponding stimuli).

conditions, the audio tracks with added noise were recombined with differing video tracks to create incongruent audiovisual stimuli in three conditions: (a) fully crossed, (b) incongruent lip rounding, and (c) incongruent height. In the fully crossed condition, vowel pairs differed in both height and rounding. In the incongruent lip-rounding and incongruent height conditions, vowel pairs differed in rounding only or in height only, respectively. This resulted in 328 stimuli of 1 s each (8 video vowel tokens + 8 audio vowel tokens × 5 noise levels + 16 incongruent audiovisual vowel stimuli × 5 noise levels × 3 conditions + 8 congruent audiovisual vowel stimuli × 5 noise levels). Thus, two different stimuli of each type were presented per condition in the control conditions (video, audio, and congruent audiovisual), and four different stimuli of each type were presented per condition in the incongruent experimental conditions.

Experimental procedure. An identification task was carried out in a sound-attenuated booth over headphones. Each participant was tested on the full set of control and experimental stimuli. Stimuli were presented and responses were collected using E-Prime Version 2.0 software (Psychology Software Tools) via a MacBook (aluminum unibody, Spring 2008 edition) running Windows XP SP2 via Boot Camp. The participants were seated facing (and about 70 cm away from) a 13.3-in. flat-panel LED display (resolution 1280 pixels × 800 pixels, vertical angle of view 18°, horizontal angle of view 26°). Participants wore Sennheiser HD 600 headphones that were directly connected to the MacBook sound card output.

The actual data collection was preceded by a short introduction with task instruction and symbol explanation (we did this to familiarize the participants with the possible responses and accompanying keys). The participants were informed that auditory, visual, and audiovisual stimuli were to be presented. The test instruction was to continuously look at the screen and to indicate by keypress what was perceived.

The test consisted of two blocks of approximately 15 min, with a short break in between. The stimuli were presented with all conditions and all stimuli randomized over both blocks. For each trial, the participant could start the presentation of the target stimulus by keypress. A fixation-cross appeared in the middle of the screen for 1 s, after which the stimulus was presented. In the audio-only condition, the screen was black. After presentation of the stimulus, the response alternatives were shown on the monitor. The possible answers consisted of all rounded and unrounded Dutch high- and mid-high–front vowels: /y, Y, ø, i, I, e/ plus the vowels /u, o, a/. These vowels were indicated on the screen with the grapheme that is normally written in Dutch with a common Dutch word to clarify the intended vowel sound. No limitation was imposed on response time.

Methodology of analysis. Perceptual confusions were measured, and confusion matrices were formed to depict patterns of perceptual change. However, in order to determine the significance of perceptual change, the experimenter must quantify the data differently, which we did by using error rates, as described below. Error rates were calculated for each experimental condition (c) by subtracting the accuracy (acc; the mean correct responses) from the highest possible error score of 1 (multiplied by 100 to obtain percentages), where acc was calculated as

$$acc(c) = \sum_{pp=1}^{Npp} \frac{N_{CORRECT(pp,c)}}{N_{TRIALS(c)}}, \qquad (1)$$

where $N_{TRIALS}(c)$ was the number of trials for condition c and $N_{CORRECT}(pp,c)$ was the number of correct responses for participant pp in condition c. We used either the visual or the auditory stimulus as a truth reference in order to determine $N_{CORRECT}$. As a means to determine the interaction effects in the audiovisually congruent conditions, we predicted error rates for multisensory responses (ε_p)

from the accuracy scores for the auditory-only and visual-only conditions as

$$\epsilon_p = 100 * \{1 - [acc(A) + acc(V) - acc(A)acc(V)]\}, \quad (2)$$

where acc(A) was the accuracy score for the audio-only condition and acc(V) was the accuracy score for the visual-only condition, and acc(A) × acc(V) was the probability that both were correct. This way, we omitted effects of statistical facilitation and only determined the possible effects of multisensory interaction.

We used a second measure, relative transmitted information score (T_{REL}), to analyze the availability of speech features in different noise conditions (for an overview and explanation, see van Son, 1994). T_{REL} was the ratio between the transmitted information, T, and the maximum rate of transmission, T_{MAX}, in percentages, such as

$$T_{REL} = 100 * \frac{T}{T_{MAX}}, \quad (3)$$

where

$$T_{MAX} = H_{STIM} + H_{RESP} \quad (4)$$

and

$$T = T_{MAX} - H_{CM}. \quad (5)$$

H_{STIM} and H_{RESP} were mean logarithmic products (entropies) for stimulus and response, respectively, and H_{CM} was the entropy of the confusion matrix, calculated by

$$H_{CM} = -\sum_{i,j} p(i,j) * \log_2 p(i,j), \quad (6)$$

where p(i, j) was the probability of observing response j for stimulus i in a two-dimensional vector or confusion matrix (H_{CM}) and was replaced by either p(i) (H_{STIM}) or p(j) (H_{RESP}) for a one-dimensional vector.

T_{REL} was calculated per feature; the analysis was performed on matrices representing either rounded and unrounded stimuli and responses, or high and mid-high stimuli and responses. The relative rate of transmission represented the ratio of the responses that can be predicted from the stimuli (Miller & Nicely, 1955).

Results

Complementarity in Congruent Audiovisual Vowels

Table 2 shows the confusion matrices aggregated over all noise levels for the congruent conditions. Note that the visual-only [Y] was more likely to be perceived as /y/ than as /Y/. All other single-channel stimuli were perceived as mostly correct. Error rates (see Figure 2) and transmitted information scores (see Figure 3) were calculated for every noise condition separately. Also, the multisensory error rates as predicted from the auditory and visual error rates are presented. Figure 2 shows the error rates for the audio-only (filled triangles), video-only (filled squares), audiovisual congruent (filled circles), and audiovisual as predicted (ε_p; open circles) conditions as a function of noise level. Vowel discrimination benefited from combined audiovisual input, which is reflected in slightly lower error rates in the congruent audiovisual condition than ε_p, the multisensory error rates as predicted from the auditory and visual error rates (Friedman's test, $\chi^2 = 2.67$, $p_{\text{one-sided}} = .051$). A post hoc comparison showed that the difference was significant for the SNR levels –6 dB, –12 dB, and –18 dB (pairwise Wilcoxon, $p_{\text{one-sided}} < .05$, adjusted for Bonferroni correction).

Visual Influence in Incongruent Audiovisual Vowels

Because the responses to incongruent stimuli can be evaluated with respect to the audio as well as the video input, we calculated two error rates for each incongruent condition. The left and right panels of Figure 4 show the error rates with regard to the auditory and visual parts of the input, respectively. The error rates for the audiovisual congruent condition are the same in both panels because the visual and auditory stimuli were the same in this condition. The figure shows that both the auditory and the visual error rates are higher in the three incongruent conditions (open symbols) than in the congruent condition (filled symbols). In all conditions, the auditory perception deteriorates with increasing noise level, which is reflected by upward slopes. In contrast, the visual perception improves with increasing noise, reflected by a similar, but inverse and less profound, pattern with regard to the visual error rates.

Table 3 shows the results of Friedman's test (a) when the visual error rate of an incongruent condition was compared with the congruent condition or the visual-only condition and (b) when the auditory error rate of an incongruent condition was compared with the congruent condition or the audio-only condition. Table 3 also shows the levels for which the post hoc Wilcoxon test is significant (after correction for Bonferroni).

For all incongruent conditions, the overall auditory and the overall visual error rates are significantly different from the four reference levels (Friedman, $p < .001$, post hoc Wilcoxon's test, adjusted $p < .05$ for all comparisons except for visual error rate for incongruent lip-rounding compared with visual-only for –6 dB, –12 dB, and –18 dB). For all but one of the conditions, the error rates in the experimental condition are significantly higher than the reference levels; namely, the overall auditory

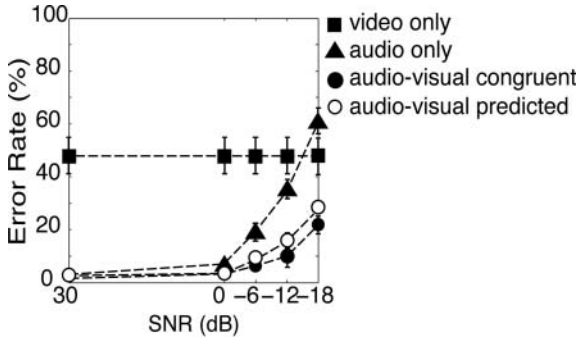
Table 2. Confusion matrices of the results of the experimental control conditions.

(A)	Stimulus audio only				(B)	Stimulus video only			
	[i]	[e]	[y]	[Y]		[i]	[e]	[y]	[Y]
Vowel	Response (%)				Vowel	Response (%)			
/i/	79.4	1.3	10.0		/i/	59.4	12.5	3.2	6.3
/I/	5.6	0.6	3.1	10.6	/I/	25.0	15.6		
/e/	1.3	77.5	0.6	13.1	/e/	3.1	56.3	3.2	3.1
/y/	10.0		76.9		/y/	3.1	6.3	67.7	40.6
/ø/		15.6		5.0	/ø/		3.1	12.9	12.5
/Y/	3.1	4.4	8.1	70.0	/Y/	6.3		9.7	25.0
/a/					/a/				
/o/		0.6		0.6	/o/			3.2	
/u/	0.6		1.3	0.6	/u/	3.1	6.3		12.5

(C)	Stimulus audiovisual congruent			
	[i]	[e]	[y]	[Y]
Vowel	Response (%)			
/i/	91.3	0.6	1.3	
/I/	6.9	2.5		0.6
/e/		96.9		
/y/	1.3		90.0	0.6
/ø/			1.3	11.3
/Y/	0.6		5.0	87.5
/a/				
/o/				
/u/			2.5	

Note. Confusion matrices in percentages (%) are rounded to 1 decimal digit. The percentages are aggregated over all presentations and noise levels for auditory stimuli (Panel A), visual stimuli (Panel B), and audiovisually congruent stimuli (Panel C). The columns and the rows represent the presented stimuli and the responses, respectively. Each cell shows the percentage of the aggregated number of times that a response was given at a stimulus presentation. The outlined cells are the responses that are congruent with either the auditory (rectangle) or the visual (oval) stimulus input, respectively.

Figure 2. Error rates for single channel, audiovisual congruent, and audiovisual predicted vowel stimuli. The depicted error rates are averaged across all listeners and are shown in percentages as a function of decreasing signal-to-noise ratio (SNR; increasing level of the steady low-pass filtered noise). The error bars show standard errors; standard errors smaller than approximately 2.5% are not visible.



error rate in the incongruent height (and, thus, congruent lip-rounding) condition is significantly lower than the audio-only error rate.

The auditory error rates in the incongruent lip-rounding condition and the incongruent lip-rounding and height condition are significantly different from the auditory error rates in both the audiovisual congruent and the audio-only conditions for the 0 dB, –6 dB, –12 dB, and –18 dB SNR levels ($p < .01$). The auditory error rates in the incongruent height condition are significantly higher than the audiovisual congruent error rates for the SNR of –18 dB ($p < .05$) and are significantly lower than the audio-only error rates for the SNR of –18 dB ($p < .05$).

Transmitted Information Scores

The transmitted information scores provide more detailed insight into the error rates, as they show what part of the information was or was not available when analyzed for different features. Figure 3 shows the transmitted information scores for lip-rounding (left panel) and height (right panel) for the audio-only, video-only, and audiovisual congruent conditions. Highest profit from visual input was in noise; the audiovisually transmitted information for lip-rounding is significantly higher than the auditorily or visually transmitted lip-rounding information for SNRs of –6 dB, –12 dB, and –18 dB (Friedman $\chi^2 = 42$ and 23, respectively, $p < .001$; Wilcoxon, adjusted $p < .05$). Furthermore the lip-rounding is better transmitted visually than auditorily at SNRs of –12 dB and –18 dB (Friedman $\chi^2 = 7$, $p < .01$; Wilcoxon, adjusted $p < .001$). The height information is better transmitted auditorily and audiovisually than visually for all SNR levels (Friedman $\chi^2 = 59$ and 66, respectively, $p < .001$; Wilcoxon, adjusted $p < .05$).

McGurk Effect in Incongruent Audiovisual Vowels

In incongruent conditions, fusions of features were expected to occur—namely, features from the visual and auditory input are recombined into a perceived vowel that was not presented in either one of the channels. We originally expected fused percepts that combine the auditorily salient height feature and the visually salient rounding feature. In the present study, apart from these expected fusions, unexpected ones were also found, as seen in the confusion matrices aggregated over all noise levels (see Table 4), where the shaded numbers represent the expected fusions. All fusions that seem to be a trend in

Figure 3. Transmitted information in percentages shown as a function of decreasing SNR (i.e., increasing noise level) for the three control conditions. The left panel denotes the transmitted information for lip-rounding, and the right panel denotes the transmitted information for height. The error bars show standard errors; standard errors smaller than approximately 2.5% are not visible.

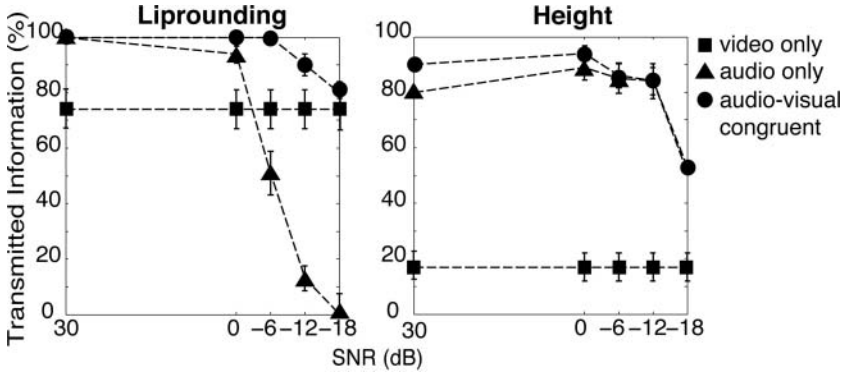
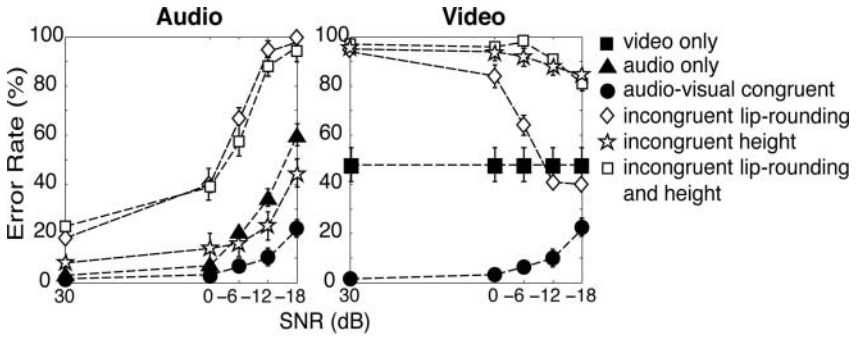


Figure 4. Error rates for audiovisually incongruent presented vowel stimuli (open symbols) as well as the reference conditions (audiovisual congruent and single channel; filled symbols). The depicted error rates are averaged across all listeners and shown in percentages as a function of decreasing SNR (i.e., increasing level of the SLN). The left panel shows the error rates with regard to the auditory stimulus, and the right panel shows the error rates with regard to the visual stimulus. The error bars show standard errors; standard errors smaller than approximately 2.5% are not visible.



the data are reported in this section, and the unexpected findings (that are a trend) are explained in the Discussion section. Furthermore, the effect of noise on the number of fusions is analyzed. An increased reliance on visual information in audiovisual perception was induced by adding noise to the auditory channel. In order to quantify this effect, we present the major fusions; these are the fusions that occur most often per category, whether they were predicted or not. Namely, if reliance on visual cues as a result of noise in the auditory domain leads to increased audiovisual integration, a significant increase in number of fusions is expected. The major fusions are plotted in Figure 5 as a function of noise. Later in the text of this article, test results for the increase (or decrease) of these fused responses is reported (see end of this section).

Table 4A shows the responses in the fully crossed condition. In this condition, fusions occurred when the vowels [i^A, e^A, y^A, Y^A] were presented with the vowels [Y^V, y^V, e^V, i^V] (where superscript "A" or "V" denotes that the phoneme was presented through the auditory {A} or visual {V} channel). Although we expected to find the fused responses /y, Y, i, e/, respectively, the observed fusions led predominantly to perceived /y, ø, i, I/ instead. The peak of the observed fusions was found at a SNR of –18 dB for [i^A] with [Y^V] and at SNR of –12 dB for the other three stimulus pairs.

Table 3. Significance tests on error rates of incongruent conditions compared with control conditions.

	Auditory error rates		Visual error rates	
Incongruent with regard to . . .	**AV-congruent**	**Audio**	**AV-congruent**	**Video**
Height				
Friedman χ^2	13***	13***	80***	57***
Significant for SNR levels (dB)	–18*	–18*	All**	All**
Lip-rounding				
Friedman χ^2	63***	58***	70***	12***
Significant for SNR levels (dB)	0, –6, –12, –18**	0, –6, –12, –18**	All**	30, 0*
Lip-rounding and height				
Friedman χ^2	65***	56***	59***	80***
Significant for SNR levels (dB)	0, –6, –12, –18**	0, –6, –12, –18**	All***	All***

Note. Summary of the statistical analysis for different experimental (incongruent) conditions. The conditions are shown in the rows, and the comparisons are shown in the columns. First, a Friedman test and, subsequently, a post hoc Wilcoxon test were applied. For the Friedman test, the χ^2 value is given, and for the Wilcoxon test, the noise levels at which the results were significant are given. AV = audiovisual; SNR = signal-to-noise ratio.

*$p < .05$. **$p < .01$. ***$p < .001$. All p values adjusted for Bonferroni correction.

Table 4. Confusion matrices of incongruent conditions.

(A) Incongruent lip-rounding and height

Audio	[i] UH	[e] UM	[y] RH	[Y] RM
Video	[Y] RM	[y] RH	[e] UM	[i] UH
Vowel	Response (%)			
/i/	58.7	0.3	34.7	3
/I/	0.7	1	13	53.3
/e/		37.3	9	10
/y/	30.7	1.3	38	
/ø/	0.3	55.3		2
/Y/	9.7	3.7	5	31.7
/a/				
/o/		1		
/u/			0.3	

(B) Incongruent lip-rounding

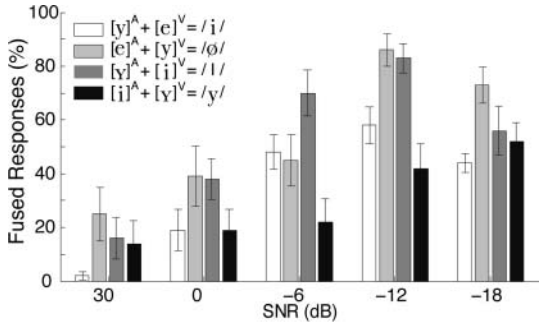
Audio	[i] UH	[e] UM	[y] RH	[Y] RM
Video	[y] RH	[Y] RM	[i] UH	[e] UM
Vowel	Response (%)			
/i/	45.3		48	0.3
/I/	1	0.7	8.7	25.7
/e/		34		40
/y/	43.3	0.7	39	
/ø/	1.3	52.7	0.7	4.3
/Y/	4	11	3.7	29.3
/a/		0.3		
/o/		0.7		
/u/	5			0.3

(C) Incongruent height

Audio	[i] UH	[e] UM	[y] RH	[Y] RM
Video	[e] UM	[i] UH	[Y] RM	[y] RH
Vowel	Response (%)			
/i/	80.3	1.7	1.7	
/I/	15.7	3		0.7
/e/	4	92		
/y/		0.3	74.7	0.3
/ø/		2.3	1.3	10.3
/Y/		0.7	22	88
/a/				
/o/				
/u/			0.3	0.7

Note. Confusion matrices in percentages (%) are rounded to 1 decimal digit. The percentages are calculated from the aggregate of responses to all presentations and noise levels for the incongruent conditions (A) lip-rounding and height (fully crossed), (B) lip-rounding, and (C) height. The columns and the rows represent the audiovisually presented stimuli and the responses, respectively. A coded description of the vowel features of the presented vowels is given in the top row (U = unrounded, R = rounded, H = high, M = mid-high). Each cell shows the percentage of the aggregated number of times that a response was given at a specific audiovisual stimulus presentation. The outlined cells are the responses that are congruent with either the auditory (rectangle) or the visual (oval) stimulus input. The shaded cells are the expected fusion responses.

Figure 5. Percentage of fused responses for the four different audiovisual vowel pairs in the crossed condition, where lip-rounding and height were both presented incongruently. The caption gives the auditory stimulus (square brackets with superscript "A"), the visual stimulus (square brackets with superscript "V"), and the fusion target (within slashes). Fusion targets are the major (not the predicted) fusions. The error bars show standard errors.



Next, Table 4B shows the responses for the lip-rounding incongruent condition. We expected increased visual responses because the auditory and visual height information are combined with the visually salient lip-rounding feature. Next to this expected result, we found that [Y^A] presented with [e^V] was sometimes perceived as /I/ (ranging from 9% at 30 dB SNR to 40% at –6 dB SNR). Also, [e^A] presented with [Y^V] was sometimes perceived as /ø/ (ranging from 34% at 30 dB SNR to 70% at –12 dB SNR).

Finally, Table 4C shows the responses for the incongruent height condition. We expected increased auditory responses because the auditory and visual rounding information are combined with the auditorily salient height feature. Next to this expected result, we found an increase in /I/ responses when [i^A] was presented with [e^V] (ranging from 0% at 30 dB SNR to 56% at –18 dB SNR).

The major fusions are shown in Figure 5 as a function of noise. For each crossed condition, a separate bar gives the percentage of fusions ranging from 1.5% fused responses for [y^A] with [e^V] in clean speech to 86% fused responses for [e^A] with [y^v] in –12dB. For all pairs, the percentage of fused responses increases with increasing noise up to SNR of –12 dB. For all pairs but [i^A] with [Y^V], the increased percentage of fused responses is turned around with SNR of –18 dB. A Friedman test with noise level as factor revealed that the amount of fused responses is significantly different for different noise levels: Friedman $\chi^2(4) = 202, p < .001$. A post hoc Wilcoxon analysis revealed that the amount of fusions significantly changed for each increase in noise. The amount of fusions increased for SNRs of 0 dB, –6 dB, and –12 dB and decreased for SNR of –18 dB (adjusted $p < .01$ for all comparisons).

Discussion

In the present study, we used congruent and incongruent Dutch front vowels as audiovisual stimuli, presented in steady low-pass filtered noise, to investigate to what extent visual cues influence the perception of vowels. The noise, by degrading the auditory input and forcing the participants to rely more on the visual input, served the purpose of producing robust perceptual interactions between the audio and visual cues.

Robert-Ribes et al. (1998) showed, for the case of vowels, that visual and auditory features are complementary (see also Summerfield, 1987)—namely, the feature whose auditory discrimination is hardest can be perceived better through vision, and vice versa. When the information is incongruent, the auditory and visual features were expected to interact in a way that can be explained by the ease of perception in either of the two channels. For incongruent stimuli, this would yield perceived vowels that combined the most salient auditory cue with the most salient visual cue. This would, in turn, lead to fusions of vowel features, similar to the McGurk effect previously observed with consonants.

Complementarity in Congruent Audiovisual Vowels

We reproduced the findings of Robert-Ribes et al. (1998) for Dutch vowels. Our results showed complementarity of the features in the auditory and visual channels; the transmitted information for lip-rounding, for example, was higher in the congruent audiovisual condition than the transmitted information for lip-rounding in the audio-only or video-only condition (see Figure 3). Also, for low SNR, the perception of congruently presented audiovisual vowels was better than the score that was predicted on the basis of vowels perceived through either of the single channels (see Figure 2).

Visual Influence in Incongruent Audiovisual Vowels

The main interest of the present study was the perception of audiovisually incongruent vowels. Because vowels are shown to contribute significantly to intelligibility of speech (Kewley-Port et al., 2007), correct perception of vowels can be decisive for speech understanding. Yet, this can be disrupted as a result of misalignment of the auditory and visual signals—for example, in modern audiovisual communication devices. Until now, research on audiovisual incongruency has focused on consonants; this research needs to be extended to vowels.

In this study, we showed that the auditory processing of vowels was influenced by incongruent visual information

that was reflected by an increase in auditory error rates in comparison to the audiovisual congruent condition (see Figure 4). The increased auditory error rate was highest for both conditions when the auditory stimulus was presented with incongruent lip-rounding; however, incongruently presented height also led to a change in the response distributions. Thus, apart from the beneficial influence that congruent visual information has on the perception of speech (Başkent & Bazo, 2011)—and, more specifically, vowels (in addition to this study, see also Robert-Ribes et al., 1998)—incongruent visual vowel information is disadvantageous for the correct perception of vowels. Even when the visual input is not very salient (i.e., height), incongruent presentation can disrupt the perceptual process, especially when the auditory signal is less well transmitted. If processing speed in audiovisual devices can be improved by passing half the auditory information, one can think of special conditions where ignoring the visually salient lip-rounding information in the audio channel of technical devices would improve the alignment by improving the processing speed. This could aid the correct perception of vowels and, hence, speech, as the information is transmitted through the channel through which it is saliently perceived.

McGurk Effect in Vowels With Incongruent Lip-Rounding

For the incongruent conditions where both visual and auditory error rates were higher than the audiovisual congruent error rates, the perceived vowel was neither the auditorily presented one nor the visually presented one. This was the case in both conditions with incongruent lip-rounding. The confusion matrices for those conditions showed fusions of vowel features (known as the *McGurk effect*). As was hypothesized, the fusions consisted mainly of vowels in which the height of the auditory vowel was combined with the rounding of the visual vowel (see shaded cells in Table 4). Exceptions to this were the following: In the incongruent lip-rounding and height condition, [Y^A] that was presented with [i^V] was perceived as /I/, and [e^A] that was presented with [y^V] was perceived as /ø/. Similarly, the incongruent lip-rounding condition showed a recombination of [Y^A] that was presented with [e^V] into /I/ percepts and [e^A] that was presented with [Y^V] into /ø/ percepts. Although we present them as exceptions, the responses can be interpreted as natural fusions. As explained in the Appendix, the vowel [Y] was used instead of [ø] because [Y] belongs to the same viseme category as [y]. Although both [ø] and [Y] are called "mid-high vowels," their first formant frequencies (F1s) are not identical (Adank, Hout, & Smits, 2004). F1 of [Y] is more similar to the F1 of [I] than to the F1 of [i] or [e]. Also, the F1 of [e] is more similar to the F1 of [ø] than to the F1 of [Y] or [y]. Therefore, the results are not intrinsically different from McGurk-like fusions, especially considering the fact that height is most salient in the auditory channel. Namely, an audiovisual vowel is perceived with the rounding of the visually presented vowel and with the F1 closest to the auditorily presented vowel.

McGurk Effect in Vowels With Incongruent Height

Contrary to our expectations, we also found significant visual influence when height was presented incongruently in the auditory and visual channels. Height is not a very visible feature because tongue placement is hidden behind lip articulation. Therefore, we expected results similar to those of the congruent stimuli—that is, [Y^A] presented with [y^V] would then lead to the auditory height perception of /Y/ and the visual rounding perception of /y/, resulting in a perceived /Y/. Indeed, the visual influence of congruent lip-rounding was additive or complementary; auditory identification improved with regard to the audio-only condition. However, next to this positive influence, we also found a detrimental influence; both auditory and visual identification degraded (i.e., resulted in higher error rates) with regard to the audiovisual congruent condition, which implies that neither the visual nor the auditory input was effectively perceived.

The confusion matrices show that the detrimental effect in both modalities was due to two effects of nonnormal perception/fusions (see Table 4C). First, [y^A] that was presented with [Y^V] led to the perception of either /y/ or /Y/, where we expected a congruency effect leading to predominantly /y/ responses. The perception of /Y/ combined the auditory and visual perception of lip-rounding with the visual height despite the fact that the visual height was less well transmitted visually than auditorily at all SNR levels (see Figure 3). Second, an increase in the number of /I/ perceptions was found when [i^A] was presented with [e^V]. This was not a purely auditory effect, as auditory [i] that was presented on its own did not often result in /I/ percepts (see Table 2A). It must be noted, however, that also in the three control conditions, /I/ responses were given to both [i] and [e] stimuli. The effect can partly be explained as follows: [I, i, e] belong to the same viseme category of short unrounded vowels (van Son, Huiskamp, Bosman, & Smoorenburg, 1994). Adank et al. (2004) showed that the mean F1 values for those vowels (pronounced by 10 female speakers) are 442 Hz, 399 Hz, and 294 Hz for [e], [I], [i], respectively. Therefore, the perceived /I/ combines the audiovisual lip-rounding with a vowel having the height (F1) in between the height of the presented vowels [e] and [i] despite the fact that height is best transmitted auditorily at

all SNR levels. It turns out that the incongruent visual input was sometimes preferred over the more reliably transmitted auditory information (see confusions in Table 4C).

It can be concluded that in special cases, where perceptual features are crossed, fusions occur in incongruently presented vowels, similar to the McGurk effect commonly observed in consonants. Vowels are longer in duration and higher in energy than consonants, and the results show evidence that these intrinsic differences do not prevent the cognitive system from binding information from the different modalities, especially when the auditory signal is less reliable. Further research could reveal audiovisual interactions between vowels and consonants. Audiovisual interactions of long vowels and short consonants could lead to partial incongruence, the effect of which is unknown. Also, the interaction of auditory and visual streams of information for people who are hard of hearing might differ from the results found in this study and, thus, needs further investigation. Namely, longstanding hearing loss might lead to a different phonological system (e.g., a few of the participants with cochlear implants in the study conducted by Schorr, Fox, Wassenhove, and Knudsen [2005] gave [ta] responses to the three different stimuli /ka, pa, ta/, indicating that the phonological system is broadened for these participants with regard to these phonemes), which could result in interactions different from the ones found here. Insight into the interaction of auditory and visual information streams in different conditions may help provide a better understanding of the problems that people experience with misalignment of the auditory and visual channels and where the focus should be with regard to alignment.

The Influence of Saliency on the McGurk Effect

The influence of saliency on the amount of fused responses can be related to the transmitted information scores. It was shown that the amount of fused responses increases significantly for increasing noise levels up to SNR of –12 dB. The auditory transmitted information scores for height decrease gradually, with noise increasing to SNR of –12 dB, and, hence, the reliance on visual information increases; transmitted information for lip-rounding is better through the visual than through the auditory channel for SNR of –6 dB and below. Furthermore, it was shown that the amount of fused responses significantly decreases for –18 dB SNR with respect to –12 dB SNR. This can similarly be related to the steep drop in transmitted information for height—and, hence, the identifiability of the height feature. Thus, when noise increases, the reliance on visual information increases accordingly, which leads to fused responses provided that the auditory height is perceived correctly.

Conclusions

In summary, we have demonstrated that the audiovisual information leads to complementarity in congruent vowels. Furthermore, we have shown that incongruent visual input influences the perception of stimuli, although visual information alone may not be sufficient to disambiguate between vowels. Finally, we have shown that this knowledge is not always used optimally, as listeners sometimes used less salient information from one modality even when more salient information was available from the other modality. The finding that even the visually less salient height feature influences auditory identification stresses the importance of appropriate audiovisual alignment in communication devices, such as cochlear implants and/or videoconferencing tools, especially when the auditory signals are degraded and listeners rely heavily on visual cues (Champoux et al., 2009; Rouger et al., 2008). For those types of applications, the addition of visual information is of great help; however, if misaligned with auditory information, it may distort the perception of speech.

Acknowledgments

The first author was supported by Stichting Technologische Wetenschappen (STW) Grant DTF 7459 and is supported by Samenwerkingsverband Noord Nederland (SNN) Grant 221. The fourth author is supported by a Rosalind Franklin Fellowship from the University of Groningen and by VIDI Grant 016.096.397 from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw).

This work was based on the second author's minor thesis, which was completed at the Research School of Behavioral and Cognitive Neurosciences, University of Groningen, the Netherlands. We thank Maeike Kiers for helping us to record the speech material and Dörte Hessler for her advice regarding the stimuli creation. Also, we thank Rob van Son for his insightful comments on the statistical analysis.

References

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America, 116,* 1729–1738.

Başkent, D., & Bazo, D. (2011). Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. *Ear and Hearing, 32,* 582–595.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9–10), 341–345.

Champoux, F., Lepore, F., Gagneú, J., & Théoret, H. (2009). Visual stimuli can impair auditory processing in cochlear implant users. *Neuropsychologia, 47,* 17–22.

Cole, R., Yan, Y., Mak, B., Fanty, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2,* 853–856.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America, 103,* 2677–2690.

Gussenhoven, C. (1999). Dutch. In International Phonetic Association (Ed.), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet* (pp. 74–77). Cambridge, United Kingdom: Cambridge University Press.

Kewley-Port, D., Burkle, Z. T., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 122,* 2365–2375.

Ladefoged, P. (1982). English vowels. In P. Ladefoged (Ed.), *A course in phonetics* (6th ed., pp. 85–105). New York, NY: Harcourt Brace Jovanovich.

Lisker, L., & Rossi, M. (1992). Auditory and visual cueing of the [± rounded] feature of vowels. *Language and Speech, 35,* 391–417.

Massaro, D. W. (1987). Single versus multiple sources of speech information: The contribution of visible speech. In *Speech perception by ear and eye: A paradigm for psychological inquiry* (pp. 27–54). Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Multiple book review of speech perception by ear and eye: A paradigm for psychological Inquiry. *Behavioral and Brain Sciences, 12,* 741–794.

Massaro, D. W., & Cohen, M. M. (1990). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 753–771.

Massaro, D. W., & Stork, D. G. (1995). Speech recognition and sensory integration: A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist, 86,* 236–244.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audiovisual speech recognition by normal hearing adults. *The Journal of the Acoustical Society of America, 77,* 678–685.

McGurk, H., & MacDonald, J. (1976, December 23). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 77,* 338–352.

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience, 25,* 5884–5893.

Pickett, J. M. (1957). Perception of vowels heard in noises of various spectra. *The Journal of the Acoustical Society of America, 29,* 613–620.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America, 53,* 1093–1101.

Rietveld, A. C. M., & van Heuven, V. J. (2009). Productie van spraakklanken [Production of speech sounds]. In *Algemene fonetiek* [General phonetics] (pp. 55–91). Bussum, the Netherlands: Uitgeverij Coutinho.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America, 103,* 3677–3689.

Rosner, B. S., & Pickering, J. B. (1994). *Vowel perception and production*. New York, NY: Oxford University Press.

Rouger, J., Fraysse, B., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Research, 1188,* 87–99.

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 18748–18759.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 3–51). Hillsdale, NJ: Erlbaum.

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics, 35,* 244–258.

van Hout, R., Adank, P., & van Heuven, V. J. (2000). Akoestische metingen van Nederlandse klinkers in algemeen Nederlands en in Zuid-Limburg [Acoustical measures of Dutch vowels: General Dutch and South Limburg]. *Taal and Tongval, 52,* 151–162.

van Son, N., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1994). Viseme classifications of Dutch consonants and vowels. *The Journal of the Acoustical Society of America, 96,* 1341–1355.

van Son, R. J. J. H. (1994). A method to quantify the error distribution in confusion matrices. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 18,* 41–63.

Appendix. Detailed analysis and justification of the selected vowels.

The high and mid-high front vowels [i, y, e, Y] were selected because lip-rounding and height features of these vowels cross in the acoustic as well as the visual domain with no other confounding features, as explained below in detail:

1. With regard to the acoustic features, height and diphthongization were aimed to be matched in pairs of vowels. The Dutch vowels [i, y] are high vowels and [e, I, Y, ø] are mid-high vowels (Adank et al., 2004; Pols, Tromp, & Plomp, 1973; van Hout, Adank, & van Heuven, 2000). Van Hout et al. (2000) found that expert listeners judged the vowels [e] and [ø] in standard Dutch as relatively monophthongal, although they are conventionally described as diphthongs (Gussenhoven, 1999) or near-diphthongs (Rietveld & van Heuven, 2009). Therefore the vowels [i, y] and [e, I, Y, ø] make appropriate candidates for the forming of vowel pairs that are either different from or equal to one another in height.
2. With regard to the visual features, the rounded vowels [y] and [Y] belong to the viseme category of "short rounded front vowels," whereas [ø] belongs to "long rounded front vowels" (Van Son et al., 1994). The vowels [e, I, i] belong to the viseme category of "unrounded front vowels." Therefore, the vowels [y, Y] and [I, i, e] make appropriate candidates for the forming of vowel pairs that are either different from or equal to one another in rounding.
3. The crossing of features in the acoustic and visual domains was necessary for analyzing the responses to the incongruent vowel stimuli—that is, where a feature can conflict in the auditory and visual domains without other conflicting features. Using crossing of features as a criterion, it was most appropriate to use [e] and [i] as monophthongal and unrounded vowels (mid-high and high, respectively) and [Y] and [y] as monophthongal and rounded vowels (mid-high and high, respectively; see Table 1). As an example, complete crossing can now be achieved by combining the auditory vowel [e] with the visual vowel [y] (crossed on both the rounding and height features, whereas all other features are kept equal).