

Two-month-olds are sensitive to lip rounding in dynamic and static speech events

Rebecca Baier¹, William J. Idsardi^{1,2}, Jeffrey Lidz^{1,2}

¹ Department of Linguistics, University of Maryland, College Park MD, USA

² Program in Neuroscience and Cognitive Science, University of Maryland, College Park MD, USA

rbaier@umd.edu, idsardi@umd.edu, jlidz@umd.edu

Abstract

Our research replicates and extends previous work on infant audio-visual speech perception [3], [4], [5], demonstrating that two-month-old infants have knowledge of the audio-visual connection between the visual aspects of lip-posture and the pronunciation of vowels and glides. Using the intermodal preferential looking paradigm [6] the infants display a clear ability to distinguish /a/-/u/, /i/-/u/ and /i/-/wi/. The infants' ability to discriminate /i/-/wi/ shows that even dynamic aspects of speech production within single syllables are salient to the infants.

Index Terms: speech perception, language acquisition, lip rounding

1. Introduction

At what point do infants know that rounded lips are characteristic attributes of certain speech sounds, such as the /u/ in 'boot'? If speech sounds are mentally represented with phonetic features such as round [1], [2], and if the features serve to link articulatory gestures with acoustic events, then we would expect even infants to display such knowledge. Pioneering work by Kuhl & Meltzoff [3], [4] and Patterson and Werker [5] provides support for this view. The previous research has shown that infants as young as two months old demonstrate knowledge of the connection between the auditory and visual components of speech. However, these previous studies only examined infants' ability to discriminate the pairs of vowels /i/-/a/ (2 months [3] and 4.5 months [1]) and /i/-/u/ (4.5 months [2]). Using the intermodal preferential looking paradigm [6] with a within subjects design, we replicate and extend this body of research by systematically investigating the visually salient feature [round] with 2-month-olds, investigating the pairs /i/-/u/, /a/-/u/ and /i/-/wi/.

Speech researchers [1], [2] have identified three different kinds of roles individual features may play in distinguishing different speech sounds in a language: contrastive, allophonic and enhancing. Using [round] as our example, we can find all three uses within the sound system of English. First, [round] is contrastive in English vowels; that is, it serves as the sole distinction between certain pairs of vowels, such as the minimal pair 'cut' /kʌt/ and 'coat' /kɔt/, whose vowels differ only in [round] (or, between 'cot' /kɒt/ and 'caught' /kɔt/ for those dialects which retain that distinction). Second, [round] is allophonic or predictable in English consonants that appear before [round] vowels. English speakers anticipate the [round] feature on the vowel, co-articulating the consonant with the vowel, for example articulating the /k/ in 'cool' with rounded lips: [k^wul]. The occurrence of [round] on consonants is predictable from the following vowel; the difference in the consonant is not meaningful by itself (though when it occurs it may be sufficient to trigger the percept of a following [round] vowel). Third, [round] is enhancing [2] with English palatal

consonants (those made with the tongue near the hard palate, such as /ʃ/ as in 'ship' and /ʒ/ as in 'genre') as it serves to make them more acoustically distinct from the similar alveolar consonants (such as /s/ in 'sip' and /z/ in 'zip'). That is, [round] serves to support another feature which makes the primary contrast (in this case [posterior]). In languages without a contrast in [posterior], such as Korean [7], the difference between [s] and [ʃ] is predictable (i.e. allophonic). In Korean this is the case, with [ʃ] used before /i/ and [s] used elsewhere, and [ʃ] is not articulated with rounded lips in Korean.

The goal of this research program is to test infants' knowledge of all of the three roles that [round] can play in languages. The current results pertain to the contrastive use of [round] in English; work currently in progress examines the allophonic and enhancing uses of [round] in English. Subsequent work will extend these results to other languages.

This report summarizes the results of three experiments on the contrastive use of [round] in English. The first experiment replicates [2] by examining contrastive lip-rounding in high vowels (/i/-/u/). The second experiment extends previous results to lip-rounding in English back vowels (/a/-/u/). Finally, the third experiment tests *dynamic* lip-rounding within a single syllable by including a round on-glide before an unround vowel (/i/-/wi/). This is the first time audio-visual knowledge of dynamic lip postures within syllables has been tested.

2. Experiment 1: /i/-/u/

Experiment 1 aimed to replicate the findings of Patterson and Werker (2003) and to provide a basis for comparison of other vowel contrasts. We asked whether infants could correctly map the phonetic properties of a vowel onto one of two faces, which differed in whether they contained visual features appropriate for the vowel. In particular, we examined the vowels /i/ and /u/, which differ in whether they involve a lip-rounding gesture, in order to determine if infants of this age can correctly identify the visual features associated with the sounds, possibly suggesting an amodal representation with modality specific correlates.

2.1. Participants

The participants consisted of 16 infants, 8 males and 8 females, ranging in age from 8.7 to 11.0 weeks (\bar{x} = 9.8 weeks, σ = 0.7 weeks). Parents were recruited through brochure mailings to families in the area with infants. Infants were not born prematurely and did not have recent ear infections or any known visual or auditory abnormalities. An additional 9 infants were excluded due to a side bias of 75% to one side (6), fussiness or looking time less than 50% (2), and researcher errors (1).

2.2. Stimulus materials

The stimuli were designed based on the descriptions in [3], [4], [5]. The test consisted of two side-by-side videos of a woman articulating sounds in synchrony (/i/ or /u/) shown on a single television. To start, the female researcher was recorded repeating the vowel sound /i/ once every two seconds for 2 minutes guided by a light flashing once every second. She tried to speak naturally, with consistent volume. The /u/ sound was recorded similarly, but watching the silent /i/ video rather than the flashing light in order to match the vowel duration more precisely. The original video was recorded on a Canon optura200MC Digital Video Camcorder.

The audio was recorded separately in a sound-treated room for 2 minutes while watching the video of the matching vowel sound. This audio was digitally recorded on a Marantz solid state recorder PMD660 using a Sennheiser head-mounted, noise-reducing microphone.

Videos were edited with Final Cut Express Version 3.5. Ten video instances were chosen for each vowel and those with the closest articulation duration were paired together. Videos that did not have extraneous head movements were chosen, and each articulation was followed by only one eye-blink. The video pairs were synchronized to the point at which the audio started for each articulation. The blinks were also synchronized by adding or deleting frames directly before or after the blink -- only when the mouth was completely closed and not moving. Cross-dissolving at the transition between video clips helped to minimize slight jumps in head location. An example frame of the resulting video is shown in Figure 1. The full stimuli (Baier_i-u.mp4) can be downloaded from <http://ling.umd.edu/labs/acquisition/?page=stimuli>.



Figure 1: *Infant visual stimuli*

Similarly, the ten best audio recordings of each vowel were chosen, checking to make sure they were similar in pitch, volume, and duration. The audio was edited in Praat [8]. We matched the sounds as closely as possible with the lengths of the original sounds from the video pairs. Average time for the /i/ sound was 0.44s, with average mouth open time of 1.36s; for /u/ the mean vowel time was 0.52s and average mouth open time was 1.32s.

The chosen audio was edited into the video in Final Cut Express and aligned with the place where the original audio had started in the videos. The 10 audio-visual pairs were continuously looped to create one 2 minute trial, with articulations once every 2 seconds. When displayed on the television, the faces were 24cm long and 20cm wide, with centers separated by 64cm. Sounds were presented from the tv speakers on each side of the screen with average intensity of 60 +/- 4 dB SPL on a Brüel & Kjær type 2240 audiometer.

2.3. Equipment and test apparatus

Stimuli were presented on a Samsung wall-mounted 51" plasma television. The testing room sized 10.3' x 8.5' had plain walls, and the experimenters stood in an adjacent control room during the experiment. Overhead lights were dimmed so that most light came directly from the television screen, and there

was not much else in the room to capture the infant's attention. To capture the infant eye movements, a Sony EVI-D100 video camera was mounted on the wall directly above the center of the television. Experimenters were able to control the pan and zoom of the camera from the control room in case the infant was fidgety.

Infants were tested one at a time, seated on a caretaker's lap, 6' from the television. The caretaker wore a visor to block their view of the video and they were asked to keep the child centered. In order to best capture infant attention, the chair was placed so that stimuli would appear at the infant's eye level.

2.4. Testing procedure

Following previous experimental designs, the procedure involved two phases: familiarization and test. The familiarization phase, which lasted 27 seconds, consisted of three silent stimuli clips. First, we presented 9s of each face (/u/ or /i/) alone, on the side that it would appear at test, then we presented 9s of both faces simultaneously, synchronized, each on the same side as at test. Before test, a colorful recentering video was presented for 4s, followed by blank screen for 2s.

The test phase consisted of two 2-minute trials. First we presented the two faces mouthing simultaneously, with one of the matching audio sounds repeating in synchronization. Between tests, the recentering video played for 4s, then we presented 2 more minutes of the same faces with the other sound. The faces stayed on the same sides of the screen throughout the familiarization and test phases. Conditions were counter-balanced for sound presented first, left-right position of the two faces, order of familiarization, and infant sex.

2.5. Scoring

Video of each infant, with a picture-in-picture window of the stimuli being shown to the infant, was recorded digitally with QuickTime software (see Figure 2). Coding of looks to the left and right was performed with SuperCoder software [9] for frame by frame analysis (29.97 frames/sec). The first author was the primary coder for 11 infants and undergraduate assistants were primary coders for 5 infants. During coding, coders were unable to hear the sound played and thus did not know which face matched the audio stimulus. A second coder coded 10% of the data. Inter-observer agreement was greater than 89%, Cohen's Kappa = 0.81.

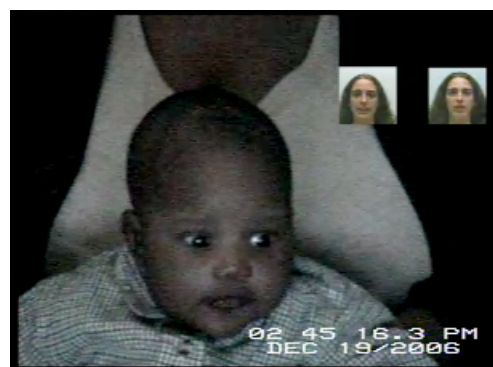


Figure 2: *Infant coding screen*

Coders marked the onset, offset and direction of each look towards one of the two faces. Custom software was then used to determine the proportion of looking time towards the matching face. This proportion is calculated by summing the looks to the matching face and dividing that sum by the total time spent looking at either face.

The familiarization phase was also coded to assess the baseline preference for either face.

2.6. Predictions

If infants are aware of the relation between the phonetic and visual properties of a vowel, then we expect them to spend a greater amount of time looking to the faces that match the audio track than to those that do not. If they are unaware of this relation, then we expect them to look equally at the two faces.

2.7. Results

Analysis of the familiarization phase revealed no preference to look toward either face (look to /u/ face = 55%, $t(14)=0.74$, $p < 0.5$, n.s). For the test phase, the infants spent 87.8% of time overall looking at either of the faces. The proportion of time infants spent looking at the matching face for the two-minute trials is shown in Figure 3. The individual subjects are plotted with dots, the mean is indicated with a green line, and the standard error of the mean with a blue line.

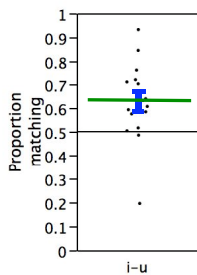


Figure 3: Matching looking time for /i/-/u/

Fourteen of the 16 infants looked more to the matching face, and overall the infants spent a significantly greater proportion of time looking at the matching face ($\bar{x} = 0.6267$, $\sigma = 0.167$, $t(15) = 3.0240$, $p = 0.0043$, one tailed).

2.8. Discussion

Infants displayed a strong preference to look at the face matching the vowel. This supports the hypothesis that infants' representations of vowel sounds include a component linking the audio and visual signals. In particular, these results show that infants are aware of both the phonetic and visual cues associated with the feature [round].

3. Experiment 2: /a/-/u/

The contrast in experiment 1 compared vowels that differed both in roundness and backness. That is, /i/ is [-round, -back] whereas /u/ is [+round, +back]. Because there is no visual cue to tongue position, we inferred that infants were sensitive to the feature [round]. However, it is possible that they are only sensitive to this feature in the context of an accompanying contrast in backness. So, the second experiment extends our results by examining vowels that are both [+back], but that contrast in vowel height and lip rounding. Assuming again that there is no visual cue to tongue position, if infants distinguish these vowels, we can infer that they are sensitive to rounding in multiple contexts.

3.1. Participants

The participants consisted of 16 infants, 8 males and 8 females, ranging in age from 8.0 to 11.1 weeks ($\bar{x} = 9.8$ weeks, $\sigma = 0.8$ weeks). Parents were recruited through brochure mailings to families in the area with infants. Infants were not born

prematurely and did not have recent ear infections or any known visual or auditory abnormalities. An additional 17 infants were excluded due to a side bias of 75% to one side (10), and fussiness/looking time less than 50% (7).

3.2. Materials and Design

The materials and design were the same as those in Experiment 1, with the substitution of /a/ as in "hot" audio and video for /i/. The /u/ materials were the same as those used in Experiment 1. Procedures for choosing the /a/ tokens and videos were identical to Experiment 1. Average time for the /a/ sound was 0.44s, with average mouth open time of 1.32s. In addition, the familiarization phase in which both faces were shown simultaneously but with no audio was increased to 18s. The full stimuli (Baier_a-u.mp4) can be downloaded from <http://ling.umd.edu/labs/acquisition/?page=stimuli>.

3.3. Predictions

If the results of Experiment 1 were due to infants' sensitivity to [round] only in the context of a contrast in backness, then we expect the proportion of looking to the matching face to be equal to chance. However, if infants have a more general sensitivity to the feature [round], then we expect infants in this experiment to look more at the matching face.

3.4. Results

Analysis of the familiarization phase revealed a significant preference to look at the /a/ face with no accompanying audio (look to /u/ face = 31%, $t(15)=-3.03$, $p < 0.009$). For the test phase, the infants spent 83.2% of time overall looking at either of the faces. The proportion of time infants spent looking at the matching face for the two-minute trials is shown in Figure 4. The individual subjects are plotted with dots, the mean is indicated with a green line, and the standard error of the mean with a blue line.

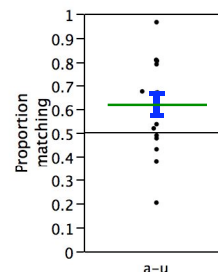


Figure 4: Matching looking time for /a/-/u/

Eleven of the 16 infants looked more to the matching face, and overall the infants spend a significantly greater proportion of time looking at the matching face ($\bar{x} = 0.6139$, $\sigma = 0.197$, $t(15) = 2.3074$, $p = 0.0179$).

3.5. Discussion

Infants displayed a strong preference to look at the face matching the vowel. This further supports the hypothesis that infants' representations of vowel sounds include a component linking the audio and visual signals. Moreover, these results show that infants are aware of both the phonetic and visual cues associated with the feature [round] independent of contrasts in tongue position.

4. Experiment 3: /i/-/wi/

Having established that infants are sensitive to both the audio and visual cues associated with the feature [round], we can

now ask whether this sensitivity is restricted to syllables consisting only of a vowel or whether infants are sensitive to these features in the context of a dynamic syllabic structure. In particular, we examine a syllable-internal sequence of a round glide (/w/) combined with a high front vowel (/i/), contrasting this with the simple vowel /i/. This is the first time syllable-internal dynamics of lip-rounding has been tested.

4.1. Participants

The participants consisted of 16 infants, 8 males and 8 females, ranging in age from 8.0 to 11.4 weeks (\bar{x} = 9.6 weeks, σ = 1.2 weeks). Parents were recruited through brochure mailings to families in the area with infants. Infants were not born prematurely and did not have recent ear infections or any known visual or auditory abnormalities. An additional 15 infants were excluded due to a side bias of 75% to one side (10), and fussiness/looking time less than 50% (2), and experimenter error (3).

4.2. Materials and Design

The materials and design were the same as those used in Experiment 1, with the substitution of /wi/ as in “we” audio and video for /u/. Procedures for choosing the /wi/ tokens and videos were identical to Experiment 1. Average time for the /wi/ sound was 0.51s, with average mouth open time of 1.36s. The /i/ materials were the same as those in Experiment 1. The full stimuli (Baier_i-wi.mp4) can be downloaded from <http://ling.umd.edu/labs/acquisition/?page=stimuli>.

4.3. Predictions

If the results of the previous experiments were due to infants' sensitivity to [round] only in the context of static syllables, then we expect the proportion of looking to the matching face to be equal to chance. However, if infants are sensitive to the feature [round] in the context of a syllable transition, then we expect infants in this experiment to look more at the matching face.

4.4. Results

Analysis of the familiarization phase revealed no preference to look at either face (look to /wi/ face 42% , $t(13) = -0.83$, $p > 0.4$, n.s.). For the test phase, the infants spent 85.5% of time overall looking at either of the faces. The proportion of time infants spent looking at the matching face for the two-minute trials is shown in Figure 5. The individual subjects are plotted with dots, the mean is indicated with a green line, and the standard error of the mean with a blue line.

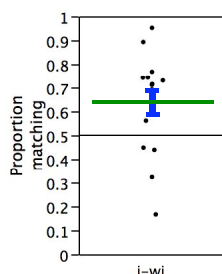


Figure 5: Matching looking time for /i/-/wi/

Twelve of the infants looked more to the matching face, and overall the infants spend a significantly greater proportion of time looking at the matching face (\bar{x} = 0.6399, σ = 0.205, $t(15) = 2.7291$, $p = 0.0078$).

4.5. Discussion

Infants displayed a strong preference to look at the face matching the syllable. This supports the hypothesis that infants' representation of the feature [round] allows it to be detected in the context of dynamic syllable transitions as well as with static syllables.

5. Pooled results

Pooling the results from the three experiments, we then analyzed the results using a full-factorial ANOVA with proportion looking time to the matching face as the dependent measure, and condition (/a-/i/, /i-/u/, /i-/wi/), order of presentation and gender as the independent variables. None of the main effects or interaction effects were significant at the $\alpha = 0.05$ level (all F ratios < 2.9, all p values > 0.10). A regression analysis with age as the predictor was also not significant ($p = 0.7942$). Thus, infants are behaving similarly across conditions and orders of presentation, and male and female infants also behave similarly in this experiment.

Another way to view the data collected throughout the three experiments is to look at the proportion of looking time to the matching face for each individual syllable (/a/, /i/, /u/, /wi/). This analysis can tell us whether there are differences in sensitivity across different syllables.

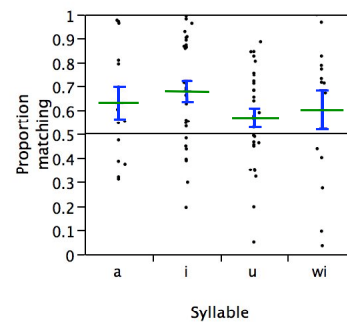


Figure 6: Matching looking time by syllable

Another ANOVA was run with proportion looking time to the matching face as the dependent measure and syllable (/a/, /i/, /u/, /wi/), order of presentation and gender as the independent variables. Only one main effect was significant the $\alpha = 0.05$ level, gender ($F = 5.6120$, $p = 0.0215$) and none of the interaction terms were significant. Figure 7 shows the distribution by gender.

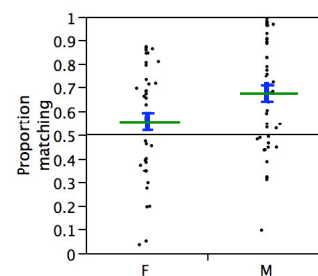


Figure 7: Matching looking time by gender

A regression analysis with age as the predictor was also not significant. However, overall the male group was slightly older than the female group, and there was no significant gender difference with any specific syllable. Moreover, both the males and the females exhibited looking times to the matching face at levels greater than chance (Female: $t(23) = 2.13$, $p < 0.05$. Male: $t(23) = 4.73$, $p < 0.0001$.) This suggests that the

better performance by the male subjects is not an important effect.

Finally, we observed some trends in the data from the familiarization phases of the experiments. As shown in Figure 8, when pairs of articulating faces are presented without audio, infants significantly prefer the [a] video over the [u] video ($t(15) = -3.0342, p = 0.0084$), and they show a slight but non-significant preference for the [u] video over the [i] video ($t(14) = 0.7418, p = 0.4705$ n.s.) and for the [i] video over the [wi] video ($t(13) = -0.8344, p = 0.4191$ n.s.).

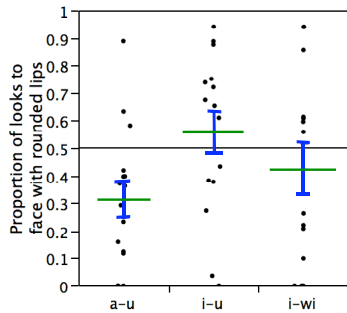


Figure 8: Proportion of looking time at the face with rounded lips by pairs of faces

However, given the wide variability in the data, an ANOVA reveals no effect for experimental condition ($F = 2.6230, p = 0.0844$ n.s.). Thus, we currently have some weak evidential trends for differential visual preferences in two month olds amongst the articulating faces, at least when there is no accompanying audio. Further research must be done to examine potential infant visual preferences.

6. General discussion

Certain phonological properties of human speech have both phonetic and visual components. In the current paper, we have examined this link for the feature [round]. Distinctive features are standardly taken as instructions to the motor system [10], [11]. Knowledge of distinctive features, then, would imply knowledge of both the phonetic and visual concomitants to the relevant articulatory gestures. We have seen that 2 month old infants have such knowledge. They are able to map the phonetic features of a syllable onto a face that is likely to have produced that syllable. This is true for the feature [round] in the context of distinctions in vowel backness as well as distinctions in vowel height. Moreover, infants are able to detect roundness in the context of dynamic syllable transitions as well as within single static segments.

These results build on the existing literature in several ways. First, prior results with 2-month old infants showed a sensitivity to the contrast between /i/ and /a/, neither of which involves the feature [round]. Second, prior results examining infants' knowledge of [round] compared /u/ and /i/ only with older (4.5 month old) infants. Third, no previous research has examined infants' sensitivity to any phonological features in dynamic syllable transitions.

Our results open the possibility of examining the feature [round] in a range of different linguistic contexts. Having established a sensitivity to this feature on glide-initial syllables, we can now ask whether infants are also sensitive to this feature in the context of consonants more generally. More narrowly, we can ask whether infants are sensitive to this feature when it is allophonic, i.e., predictable from the features of a following vowel, and when it is enhancing, i.e., providing phonetic support for a primary contrastive feature. Finally, these results also open the door for research on fused audio-

visual percepts, including phenomena such as the McGurk effect [12], in which, for example audio [b] combines with visual [g] to yield a perceived [d] which is different from both of the component stimuli. It will be especially interesting and useful to chart the development in children of the size of the fairly large integration window for audio-visual speech, which in adults runs from about 30 ms of audio lead to 170 ms of audio lag [13].

7. Acknowledgements

This research was supported in part by a grant from the National Institutes of Health (R03-DC006829). This paper has benefited from helpful discussion with Colin Phillips, David Poeppel and Norbert Hornstein. Thanks also to Shannon McDaniel for assistance in creating the stimuli and for coding, along with Cynthia Lukyanenko, Lykara Charters, and Alyssa Rodriguez. Many thanks to the undergraduate research assistants who helped run the experiment, and to the parents and children who participated in this research.

8. References

- [1] Jakobson, R., Fant, G., and Halle, M. (1951) *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.
- [2] Keyser, S.J., and Stevens, K.N. (2006) Enhancement and overlap in the speech chain. *Language*, 82, 33-63.
- [3] Kuhl, P.K., and Meltzoff, A.N. (1982). The Bimodal perception of speech in infancy. *Science*, 218, 1138-1141.
- [4] Kuhl, P.K., and Meltzoff, A.N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy: Minnesota symposia on child psychology* (Vol. 20, pp. 235-266). Hillsdale, NJ: Erlbaum.
- [5] Patterson, M.L., and Werker, J.F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6, 193-198.
- [6] Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, 15, 626-636.
- [7] Sohn, H-M. (1999) *The Korean Language*. Cambridge: Cambridge University Press.
- [8] <http://www.praat.org>
- [9] Hollich, G. (2005). Supercoder: A program for coding preferential looking (Version 1.5). [Computer Software]. West Lafayette: Purdue University.
- [10] Jakobson, R., Fant, G. and Halle, M. (1951) *Preliminaries to Speech Analysis*. Cambridge MA: MIT Press.
- [11] Lieberman, A. (1996) *Speech: A Special Code*. Cambridge MA: MIT Press.
- [12] McGurk, H. and MacDonald, J. (1976) "Hearing lips and seeing voices," *Nature*, 264, 746-748.
- [13] van Wassenhove V, Grant KW, Poeppel D. (2007) Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45, 598-607.