Attention to the eyes and mouth during audiovisual speech perception:

A comparison of looking behavior in infants and adults

Emily J. Roemer

University of Rochester

Abstract

Audiovisual speech perception plays an important role in language development, and recent eye-tracking research has examined attention to the eyes and mouth in an effort to better understand how infants use visual cues during speech. Lewkowicz and Hansen-Tift (2012) found an attentional shift from the eyes to the mouth in 6 month old infants, and back to the eyes by around 12 months, suggesting the importance of integrating visual cues from the mouth during this sensitive period in language development. In the current study, we investigate these eye-gaze patterns further by testing whether infants look more to the eyes or to the mouth during different types of speech sounds in nonsense words, including an audiovisual mismatch condition (i.e. the "McGurk effect"). Infants between 12 to 15 months of age, as well as an adult comparison group, were tested in the present eye-tracking study. Participants watched videos of three different conditions of nonsense words: words containing *b* or *p* sounds (i.e. "bapu"), words containing *g* or *k* sounds (i.e. "gaku"), and an audiovisual mismatch of the two (i.e. audio "gaku" dubbed onto visual "bapu"). While minimal differences between these conditions were found, when comparing each condition to a baseline proportion of looking time when viewing a still face, we found that infants look more to the eyes on a still face (83%), but significantly less to the eyes (and more to the mouth) during speech across conditions. Adults also look more to the eyes on a still face (78%), but only significantly more to the mouth during audiovisual mismatch conditions. The present study also provides evidence that infants listening to these nonsense words look significantly more to the mouth overall while adults look more to the eyes, suggesting infants at this age are still developing their use of audiovisual cues in speech perception.

Key words: audiovisual speech perception, McGurk effect, infants, eye-tracking

Attention to the eyes and mouth during audiovisual speech perception: A comparison of looking behavior in infants and adults

Audiovisual speech perception plays an important role in the development of social communication abilities, and this process has been extensively studied throughout the life span (Burnham & Dodd, 2004; McGurk & MacDonald, 1976; Munhall & Vatikiotis-Bateson, 2004; Pons, Lewkowicz, Soto-Faraco and Sebastian-Galles, 2009; Summerfield & McGrath, 1984). When someone speaks, people combine many sources of information to understand the utterance; visible speech information, gaze direction, facial expressions, gestures and tone of voice all contribute to a complete understanding (Munhall & Vatikiotis-Bateson, 2004). Speech perception is therefore a multimodal process, and examining the ability of infants to integrate this information is important to our understanding of language development.

Recent eye-tracking research has examined infants' attention to the eyes and mouth while listening to speech in an effort to understand how infants gain information about the speech sounds of their language. Lewkowicz and Hansen-Tift (2012) tracked eye-gaze patterns in 4 to 12 month old infants and found a shift in eye-gaze from the eyes to the mouth around 6 months of age, and back to the eyes between 10 and 12 months of age when listening to the infants' native language. This shift holds true for native language, but Lewkowicz and Hansen-Tift (2012) also found that when listening to a non-native language (Spanish), infants continue to look more at the mouth at 12 months, suggesting they are seeking out visual cues to understand the nonnative speech sounds. The present study seeks to investigate these eye-gaze patterns further by testing whether 12-15 month old infants look more at the eyes or mouth for nonsense words using English speech sounds (i.e., bapu, kogee, etc.), and also explore gaze patterns during

an audiovisual mismatch condition combining these speech sounds (i.e., audio gaku "dubbed" onto visual bapu).

The shift seen in infants' visual attention to the mouth during the time period between 6 and 12 months has interesting implications for language development; it suggests that infants are gaining information about how to produce the sounds of their native language through imitation (Lewkowicz & Hansen-Tift, 2012). This type of learning through imitation of audiovisual cues at this time in development makes sense given the concurrent onset of babbling, and the idea that social learning is vital to the normal development of language (Kuhl, 2007). Interestingly, the shift in looking to the mouth between 6 and 12 months also coincides with the time period of perceptual narrowing seen across several domains. Research has established a narrowing of phoneme perception between 6 and 12 months; at 6 months infants can discriminate non-native phonemes, but by 10-12 months they can only discriminate phonemes of their native language (Kuhl, Stevens, Hayashi, Deguchi, Kiritani & Iverson, 2006). A similar narrowing effect has been shown in face perception between 6 and 9 months for both other-species faces (Pascalis, de Haan & Nelson, 2002) and other-race faces (Kelly, Quinn, Slater, Lee, Ge & Pascalis, 2007). Pons et al. (2009) have extended this perceptual narrowing effect to multisensory speech, with findings suggesting that infants become perceptually tuned to audiovisual correspondences of their native language between 6 and 11 months of age.

With perceptual narrowing showing up across so many domains, it is interesting to consider Lewkowicz and Hansen-Tift's (2012) finding of an attentional shift to the mouth in the context of audio-visual integration during speech perception. The importance of redundant audiovisual speech cues becomes obvious when looking at the effects of creating a mismatch between auditory and visual cues, as can be seen with the well-known "McGurk effect". When

auditory and visual cues do not match (i.e., an auditory "ga" is paired with a visual "ba"), people unconsciously try to integrate these cues into a perception that makes sense (i.e., a perceived "da"). Even though adults tend to look more at the eyes during speech perception (Lewkowicz & Hansen-Tift, 2012) and so would seemingly not notice the conflicting auditory and visual cues, the McGurk effect shows up strongly in adults (McGurk & MacDonald, 1976; Summerfield & McGrath, 1984; Buchan & Munhall, 2012). This may be due to adults' ability to use cues from their peripheral vision, which would allow them to integrate visual information from the mouth while looking to the eyes for social information. The McGurk effect has been studied in infants as well, with findings that 5 month old infants are influenced by the McGurk effect (Burnham & Dodd, 2004; Rosenblum, Schmuckler & Johnson, 1997) and that event-related potentials (ERP's) are associated with looking time to the mouth during presentation of a McGurk illusion between 6 and 9 months (Kushnerenko, Tomalski, Ballieux, Ribeiro, Potton, Axelsson, Murphy & Moore, 2013). These findings suggest that infants already have the ability to integrate auditory and visual speech information at a young age.

While attention to the eyes and mouth has been investigated across the first year of life during native and non-native speech perception (Lewkowicz & Hansen-Tift, 2012), and infants have been shown to integrate auditory and visual speech information (Burnham & Dodd, 2004; Kushnerenko et al., 2013; Rosenblum et al., 1997), looking behavior during different types of speech sounds has not been thoroughly investigated. The present study examines attention to the eyes and mouth during three different conditions: "visible", which include the more visually salient labial phonemes *b* and *p* (e.g., bapu, peeba); "invisible", which include the less visually salient velar phonemes *k* and *g* (e.g., gaku, keega); and "dubbed", which include words with an audiovisual mismatch (e.g., visual bapu spliced onto auditory gaku). We investigated this in 12-

15 month old infants, as this is the age that Lewkowicz and Hansen-Tift (2012) found an attentional shift back to the eyes for native language but not non-native language. We also conducted the study with adults as a comparison group, since adults have been shown to look more towards the eyes even for non-native language (Lewkowicz and Hansen-Tift, 2012). We hypothesized that infants would look more to the mouth during "visible" conditions than "invisible" conditions given that there is more visually apparent speech information in labial sounds than velar sounds. We also hypothesized that if infants notice the audiovisual mismatch in the "dubbed" conditions, they would look more to the mouth to examine the confusing sounds more closely. Finally, we hypothesized that infants would look less to the eyes than adults, as they are still developing their audiovisual integration abilities and use of social cues from the eyes during speech perception.

Methods

Participants

Participants included twenty infants age 12 to 15 months (9 females and 11 males; mean age = 13 months 3 days; SD = 19 days) and fourteen adults (12 females and 2 males; mean age = 20 years). Both infants and adults were assigned to one of two stimulus lists, described below. All infants were healthy with no visual or auditory impairment, born full-term (no more than 3 weeks before their expected due date), and classified by the parents as hearing at least 75% English. Participants were recruited using a database comprised of families willing to participate in studies at the Rochester Baby Lab. Families in this database were recruited using a variety of methods including mailings from several hospitals in the Rochester area, flyers posted in day cares and clinics, and booths at public events like the farmer's market. Consent was obtained from a parent/guardian of each participant, and parents were paid $10 for their participation.

Adults were undergraduate students at the University of Rochester recruited through association with the researchers. All adults were native English speakers, though two were bilingual and 8 others had significant familiarity with other languages. Consent was obtained and adult participants were not compensated for their participation.

Design

Each participant watched a five-minute video consisting of a female speaker saying two-syllable consonant-vowel-consonant-vowel nonsense words (see Figure 1). These words were categorized in three different conditions: "visible", which contained the visually salient phonemes *b* and *p* (e.g., bapu, peeba); "invisible", which contained corresponding words with the less visually salient phonemes *g* and *k* (e.g., gaku, keega); and "dubbed", in which the audio "invisible" words were spliced onto the corresponding visual "visible" words (e.g., visual bapu/ audio gaku). The speaker was video and audio recorded saying each word against a gray background with only her face and shoulders visible. For each trial, the word was spoken twice with approximately 2 seconds in between. Each clip lasts approximately 8 seconds, and tone of voice was consistent throughout. For the "dubbed" trials, each clip was edited such that the onset of each word matched between audio and video.

Two list orders were created using identical stimuli, each with 8 blocks containing a trial from each of the three conditions (24 trials total for each list, see Table 1). Each list was presented in a pseudo-random order with no block containing two corresponding words (e.g., bapu, gaku, and dubbed bapu/gaku were never in the same block), and the presentation of the three conditions randomized (e.g., dubbed-visible-invisible; visible-invisible-dubbed; invisible-visible-dubbed; etc.). Each list was created in Tobii Studio (3.1.6) with a 2-second black screen presented after each trial. An attention-grabbing video (a small circle with the face of a laughing

baby) was presented every two blocks for adults and after every block for infants. For both adults and infants, half of the participants (every other participant) watched list 1 and half watched list 2.

Procedure and Analyses

Eye-tracking data were collected using a Tobii eye tracker with a 24 inch monitor, which uses an infrared light source and camera below the flat-screen monitor to record corneal reflection data to determine gaze direction. The Tobii system uses a combination of measurements from each eye to determine where on the screen the participant is looking. Both adult and infant participants sat approximately 60 cm from the screen (infants sitting on their parent's lap), and an experimenter adjusted the height of the screen until the Tobii system picked up on the location of the participant's eyes. A five-point calibration procedure was run, and recording and stimulus presentation started once each point was calibrated correctly. Adult participants also completed a brief follow-up survey with the questions "What do you think we are studying?" and "Did you develop any patterns or strategies in understanding the speech sounds?" to see if they noticed the combined audio/video on the dubbed trials.

To determine the proportion of time participants spent fixating on the eyes and mouth, areas of interest (AOI's) for these areas were defined in Tobii Studio. The AOI for the eyes was a rectangular area extending vertically from directly above the eyebrows to the top of the cheekbones and horizontally to the edges of the speaker's face. The AOI for the mouth was a rectangular area of the same size, beginning mid-way between the nose and the upper lip and extending the same height and width as the eye-AOI. Eye-tracking data for both AOI's was recorded for the duration of each trial, and every 16 milliseconds a data point was generated for each region of interest. A "1" was generated when the participant was looking at the AOI, and a

"0" if they were looking elsewhere or if the eye tracker did not pick up their eyes at that time point. Proportions for each AOI were calculated by taking the sum for each AOI in each trial and dividing it by the sum of both AOI's (eyes and mouth) combined. This gives a proportion out of the total time looking at the eyes or the mouth, so that the two proportions add up to 1. Each trial was determined to be usable if at least 40% of the data points generated good eye-tracking, defined as Tobii validity scores of 0, 1 or 2 (on a 0-4 scale with 0 being perfect tracking) for both eyes. Fourteen infants and three adults were excluded from the analysis due to insufficient eye-tracking data or excessive fussing. Sufficient eye-tracking data was defined as 2 or more usable trials from each condition (dubbed, visible and invisible). Only usable trials (at least 40% good eye-tracking) were used in the analysis. After these exclusions, 11 adults (9 females and 2 males; mean age 20 years) and 6 infants (3 females and 3 males; mean age 12 months, 27 days) were included in the analysis. Approximately half of these (5 adults and 3 infants) watched List 1 and the rest (6 adults and 3 infants) watched List 2.

Results

List 1 and List 2 were combined for the infant data without checking for list order effects due to only having three participants with usable data in each group. A two-factor ANOVA with repeated measures on one factor was run on the adult data to check for list order effects, and no significant difference was found (P=0.1347). For all further analysis, list 1 and list 2 were combined (see Table 2 for list means). All analyses were run using the eye proportions, as eye and mouth proportions were calculated in such a way that they are reciprocals of each other (see Figure 2). A one-way ANOVA for correlated samples with three samples (dubbed, visible, invisible) was run on the infant data, and the difference between conditions bordered on significance (P=0.0506). Two-sample correlated t-tests were then run between each condition

(dubbed/invisible, dubbed/visible, visible/invisible) to check which conditions might contribute to this difference. The only significant finding was a difference between visible and invisible conditions (One-tailed $P<0.02$), with infants looking more to the mouth during the invisible condition (gaku, kogee, etc.) than the visible condition (bapu, pobee, etc.). A one-way ANOVA for correlated samples was also run on the adult data to test for differences between the three conditions, but no significant differences were found.

Although minimal differences between conditions were found, examination of the data shows an obvious preference for infants to look more to the mouth than the eyes, and adults to look more to the eyes than the mouth, across conditions (see Figure 2). Rather than assume a mean of 50% as chance, we decided to determine the base rate of looking to the eyes and mouth for both infants and adults. To do this, we examined gaze behavior during the first 2 seconds of data for each participant, during which the screen showed the speaker's still face before speech. We only took this information from the first trial for each participant, as we did not want expectations based on previous trials to influence the base rate. After calculating proportions looking to each AOI for these initial 2 seconds and taking the mean across participants, we found that infants look to the eyes at a rate of 0.83, and adults look to the eyes at a rate of 0.78. These proportions reflect looking time to the eyes divided by the sum of time looking to either the eyes or the mouth prior to the onset of speech. We then ran single sample t-tests using these base rates as a hypothetical mean, to determine whether overall proportions of looking to the eyes in each condition was significantly different from the proportion of looking to the eyes of a still face. In infants, these were significant in all conditions (Dubbed: $P<0.01$; Visible: $P<0.01$; Invisible: $P<0.001$). This suggests that infants at the age of 12 to 15 months look more to the mouth while watching a speaker say nonsense words than they do while viewing a static face. In adults, these

single sample t-tests compared to the hypothetical mean of 0.78 only yielded a significant result in the dubbed condition ($P<0.05$), although the other conditions approached significance (Visible: $P=0.0854$; Invisible: $P=0.0631$). This suggests that adults look more to the mouth while viewing an audiovisual mismatch than they do while viewing a static face.

Finally, a two-way factorial ANOVA for independent samples was run across all conditions for infants and adults, with age group as an independent factor. This yielded a significant effect of age group ($P<0.01$), showing that looking time to the eyes and mouth differed significantly between infants and adults.

Discussion

This study sought to explore looking behavior to the eyes and mouth in infants and adults while listening to different types of speech sounds, including an audiovisual mismatch condition. Lewkowicz and Hansen-Tift (2012) found a shift in looking behavior from the mouth to the eyes by around 12 months for the native language, but not for non-native languages. We explored this further by investigating looking behavior in 12-15 month olds, after the expected normal shift back to the eyes, but for different types of speech sounds. We also conducted the study with adults to explore whether infants of this age have different gaze patterns than adults for the audiovisual speech mismatch condition (the "McGurk effect"). While the McGurk effect has been shown to be present in infants as young as 5 months of age (Burnham & Dodd, 2004; Rosenblum et al., 1997), infant gaze patterns in response to this effect have not been extensively studied. We hypothesized that infants would look more to the mouth during the audiovisual mismatch than with other speech sounds to attempt to disambiguate the conflicting cues. We also predicted that infants would look more to the mouth than adults, as they still depend on

redundant audiovisual cues for sounds they have not yet mastered (Lewkowicz and Hansen-Tift, 2012).

One serious limitation for this experiment is that we are unable to determine empirically the reason for the lack of usable eye-tracking data in our infant participants. Out of 20 infants, only 6 had at least two usable trials from each condition. While it may be that the infants were simply not attentive and looked away from the screen, it appeared that many infants were engaged throughout the video but the Tobii system could not track their eyes, despite good calibration prior to the experiment. Due to this lack of good eye-tracking, we only included data from 6 infants in the final analysis, and had to collapse our two list order groups across infant participants. In examining our results, it is important to keep this limitation in mind.

Contrary to our predictions, there were minimal significant differences between conditions, with the only difference being that infants looked more to the mouth for "invisible" (e.g., gaku, kogee) sounds than "visible" (e.g., bapu, pobee) sounds. It may be that the infants did not notice the audiovisual mismatch in the "dubbed" condition, and so did not look to the mouth any differently for these than for other sounds. The reason for the infants preference for looking to the mouth more for "ga/ka" sounds than "ba/pa" sounds is not clear, though one possibility is that due to these sounds being less visually salient, infants are looking to the mouth for longer periods to see if they can gather any additional information, while for "ba/pa" sounds they can visually discriminate these from other sounds more quickly. Further testing with a larger sample and more efficient eye-tracking equipment would help to determine whether this difference remains significant and if any other differences arise.

Our additional findings support our hypothesis that infants look more to the mouth (and less to the eyes) during speech than adults. We found that after taking into account a high base

rate of looking to the eyes prior to any movement in the speech video, infants look significantly more to the mouth overall across conditions. We were intrigued to find that the base rate of looking to the eyes prior to speech was so high (83%) given that the overall proportions showed a tendency for infants to look more to the mouth overall. This suggests that infants are looking towards the eyes for social cues prior to speech, but simply look more to the mouth once speech starts in order to utilize visual information about the vocal gestures. Research has shown that infants as young as 3 months can discriminate adult gaze direction (Hood, Willen, & Driver, 1998), so it makes sense that infants at 12 months would look to the eyes more to gain social information from a still face, even though gaze direction is not informative in our experiment.

When taking into account the base rate of adults looking to the eyes prior to movement, we found that adults only look significantly more to the mouth during the audiovisual mismatch ("dubbed") condition, although the other two conditions approached significance. This suggests that adults look more to the mouth to disambiguate confusing speech sounds, which is consistent with research showing adults look more to the mouth in audiovisual speech in noisy conditions when it is challenging to understand the speech from auditory cues alone (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998).

Overall, our finding that adults look more to the eyes than infants is consistent with Lewkowicz and Hansen-Tift's findings (2012) that an attentional shift back to the eyes occurs around 12 months of age when listening to native language, but sometime later in development for sounds with which an infant is unfamiliar. Though our conditions contained speech sounds from English, we used nonsense words that the infants presumably had never heard before, as well as ambiguous "dubbed" speech sounds. While we found minimal significant differences between conditions, our research does support our hypothesis that infants at 12-15 months of age

look more to the mouth than the eyes, while adults look more to the eyes than the mouth. The small sample size and lack of usable eye-tracking data makes these findings difficult to generalize, and future research should investigate gaze patterns both for different speech sounds and across different ages to further explore these findings across development.

References

Buchan, J. N., & Munhall, K. G. (2012). The effect of a concurrent working memory task and temporal offsets on the integration of auditory and visual speech information. *Seeing and Perceiving*, *25*(1), 87.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204-220.

Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, *9*(2), 131-134.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, *18*(12), 1084-1089.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13-F21.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental science*, *10*(1), 110-120.

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., Murphy, E., & Moore, D. G. (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *European Journal of Neuroscience*, *38*(9), 3363-3369.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431-1436.

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

Munhall, K. G., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 177-188). Cambridge, MA: MIT Press.

Pascalis, O., de Haan, M., and Nelson, C.A. (2002). Is face processing species-specific during the first year of life? *Science,* 296, 1321-1323.

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences*, *106*(26), 10598-10602.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347-357.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology*, *36*(1), 51-74.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*(6), 926-940.

Table 1

List orders

List 1			List 2		
Pabu/Kagu	Peeba	Kugo	Bapu	Kogee	Pabu/Kagu
Pobee	Gaku	Bopee/Gokee	Pobee	Kogee	Pubo/Kugo
Keega	Bupa	Pubo/Kugo	Keega	Bupa/Guka	Pubo
Pobee/Kogee	Kagu	Pubo	Pobee/Kogee	Gaku	Bopee
Bapu	Gokee	Beepo/Geeko	Guka	Peeba	Beepo/Geeko
Guka	Peeba/Keega	Beepo	Bopee/Gokee	Bupa	Geeko
Bapu/Gaku	Kogee	Pabu	Bapu/Gaku	Gokee	Beepo
Bopee	Bupa/Guka	Geeko	Pabu	Peeba/Keega	Kugo

Note. Each row is one block, containing a dubbed, visible and invisible trial. Each word is repeated once.

Table 2

Mean proportions of time looking at either the eyes or the mouth

	Dubbed		Visible		Invisible	
Group	Eyes	Mouth	Eyes	Mouth	Eyes	Mouth
Adults List 1	0.7699	0.2301	0.8018	0.1982	0.7683	0.2317
Adults List 2	0.4772	0.5228	0.5082	0.4918	0.5298	0.4702
Adults combined	0.6102	0.3898	0.6417	0.3583	0.6381	0.3618
Infants List 1	0.4323	0.5677	0.4343	0.5657	0.3831	0.6169
Infants List 2	0.3673	0.6327	0.4148	0.5852	0.3244	0.6756
Infants combined	0.3998	0.6002	0.4246	0.5754	0.3538	0.6462

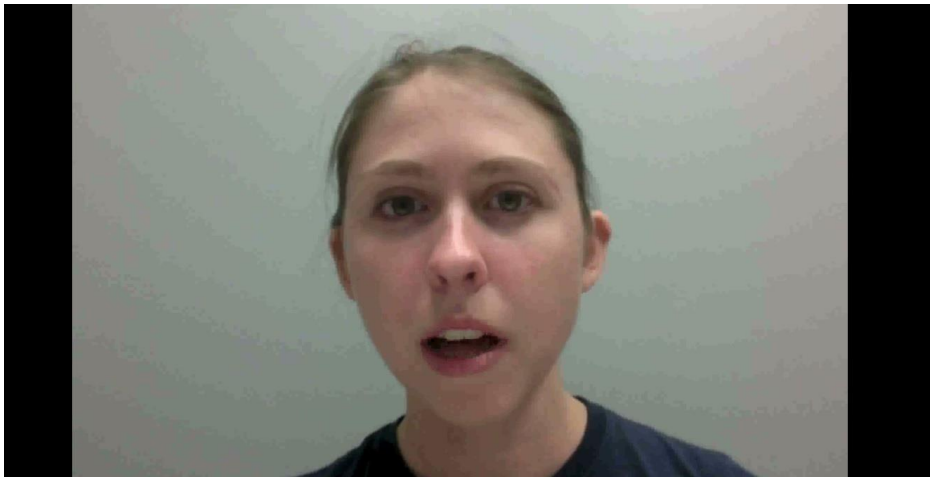

Figure 1. Video stimulus: a female speaker against a gray background. Positioning in the frame was identical for each trial.

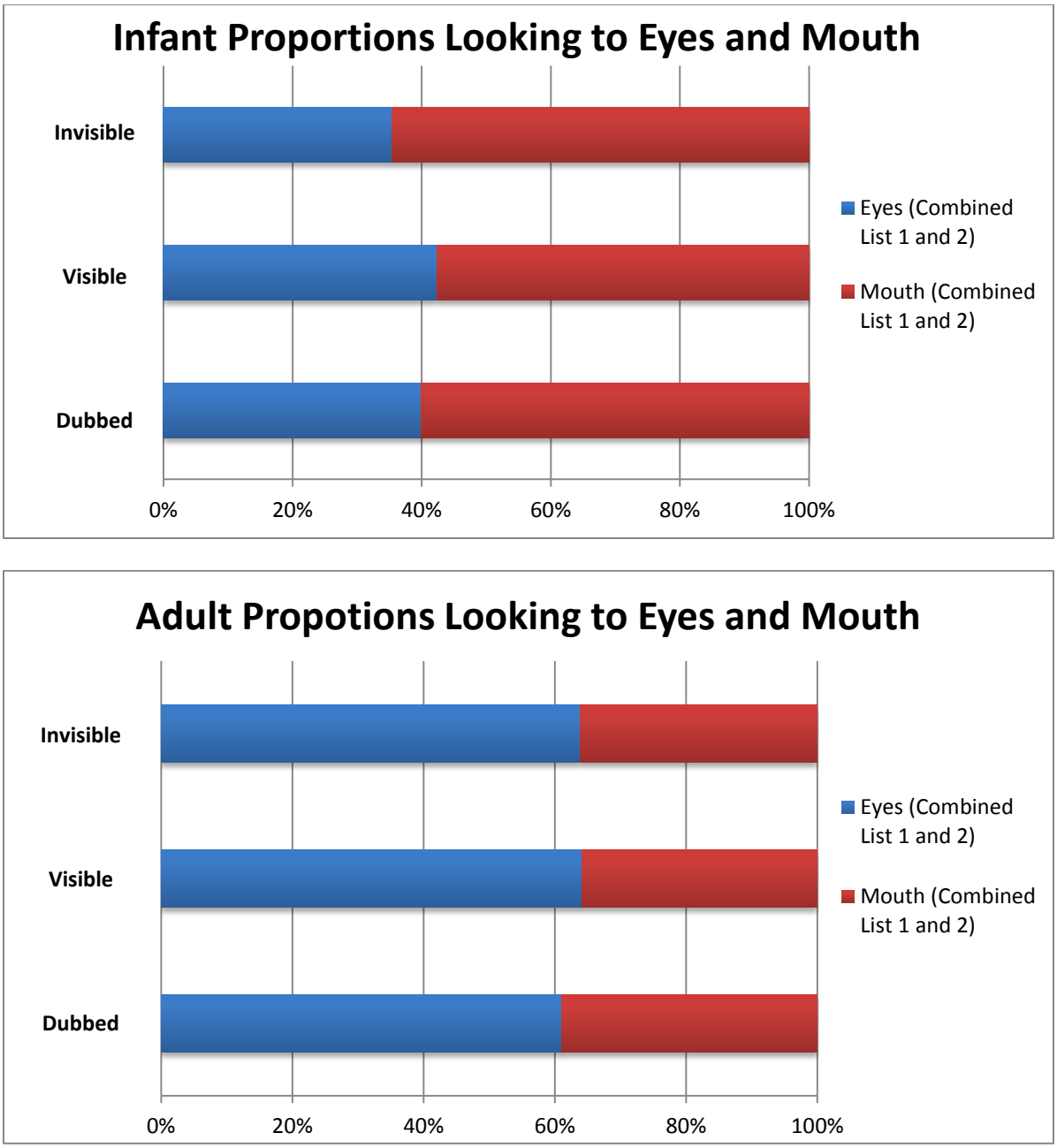

Figure 2. Mean proportions out of total time looking to either the eyes or the mouth throughout each trial for each condition.