Phonetic information in audiovisual speech is more important for adults than for infants; preliminary findings.

Martijn Baart[1], *Jean Vroomen*[2], *Kathleen E. Shaw*[3]*& Heather Bortfeld*[3,4]

[1]Basque Center on Cognition, Brain and Language, Donostia, Spain.
[2]Dept. of Cognitive Neuropsychology, Tilburg University, Tilburg, the Netherlands.
[3]Dept. of Psychology, University of Connecticut, Storrs, CT, USA.
[4]Haskins Laboratories, New haven, CT, USA.

M.Baart@bcbl.eu, J.Vroomen@uvt.nl, Kathleen.Shaw@uconn.edu, Heather.Bortfeld@uconn.edu

Abstract

Infants and adults are able to match auditory and visual speech but the cues on which they rely may differ. Here we provide an initial assessment of the relative contribution of temporal- and phonetic cues available in the AV signal. Adults (N=52) and infants (N=18) matched 2 trisyllabic speech sounds, either natural speech or SWS, with visual speech information. Adults saw two articulating faces and matched a sound to one of these, while infants were presented with the same stimuli in a preferential looking paradigm. Adults' performance was almost flawless with natural speech, but was significantly less accurate with SWS. In contrast, infants matched the sound to the articulating face, irrespective of whether it was natural speech or SWS. We propose that infants matched the AV signal based on temporal cues whereas adults relied more heavily on phonetic cues. This is in line with the idea that lipreading improves with age.

Index Terms: Phonetic correspondence, temporal correspondence, audiovisual speech, sine-wave speech

1. Introduction

Adults and infants integrate auditory and visual speech into one event [e.g., 1, 2-7], but they may rely on different cues, viz. temporal versus phonetic..

Audiovisual (hence, AV) temporal cues are derived from bimodal characteristics such as speech rate and AV onset of syllables. Both adults and infants are sensitive to these temporal dynamics as they are able to detect AV asynchrony, [e.g., 8, 9-11].

AV phonetic cues can be derived from the phonetic content in sound and vision. For instance, a listener recognizes that a bilabial closure may correspond to /m/ but not to /s/). Although infants are sensitive to phonetic information in the (AV) speech signal [3, 4, 12-15], the ability to extract phonetic content from visual speech increases with age and develops well beyond puberty [6, 16-22].

Adult AV speech integration is achieved at multiple levels [23], which has recently been corroborated by studies using sine-wave speech [24, 25]. In sine-wave speech [i.e., SWS, see 26], the natural richness of the speech signal is reduced to a few sinusoids that track the center-frequencies of the lowest formants (usually F1, F2 and F3). Critically, the temporal dynamics of the natural speech signal are retained but listeners do usually not perceive SWS as speech without explicit instruction [e.g., 26]. Perception of auditory and visual temporal order, though, is independent of whether SWS is heard as speech or not [25] whereas visual speech induced phonetic biases in auditory speech identification only occur when listeners hear the phonetic content in the SWS sounds [24, 25, 27, 28].

When presented with two simultaneous videos of a speaker talking, infants prefer to look at the speaker whose visual speech matches a speech sound they are hearing [e.g., 1, 2, 29].

Although this can be taken as evidence that infants can extract and use AV phonetic correspondence from the signal, phonetic information is not always needed in both signals for infants to detect AV correspondence. For example, infants can separate a target auditory speech passage from a distracting one based on a non-speech visual signal synchronized with the auditory speech input [i.e., an oscilloscope pattern, 30].

Here, we compared infants' and adults' detection of AV speech correspondence based on temporal and phonetic cues. We used two trilsyllabic AV pseudo-words, in which the sound was either natural speech or SWS. The rationale was that both natural speech and SWS contain temporal cues whereas phonetic cues are most prominently available in natural speech. This would imply that AV correspondence detection can be based on both crossmodal cues in natural speech, whereas AV correspondence detection for SWS is mainly driven by temporal coincidence.

Adults were tested with a forced choice matching task (i.e., which of the two faces matches the audio?) and were presented with either natural speech or SWS. For the infants we used a preferential looking procedure.

We hypothesized that adults would perform worse with SWS than with natural speech because only the latter contains for adults beneficial phonetic cues. For infants, we hypothesized that, if the ability to extract phonetic content from the AV signal indeed develops over time, the difference between AV correspondence detection for natural speech versus SWS would be smaller than in adults.

2. Methods

2.1. Participants

2.1.1 Adults

52 undergraduate students (Mean age = 19.5 years) from the University of Connecticut participated in return for course credits after giving their written informed consent. Participants were assigned to either the natural speech- (NS) or the SWS group (N = 26, 13 females in both groups).

2.1.2 Infants

18 infants in between 8 and 12 months of age participated and were randomly assigned to either the natural speech (NS) group or the SWS group (N = 9 in both groups).

2.2. Stimuli

Stimulus creation began with recording a female native speaker of Dutch (with a video-camera) pronouncing two three-syllable CV-strings that made up the pseudo-words 'kalisu' and 'mufapi' The audio was extracted, cut-off at onset and background noise was removed with the Adobe Audition 3.0 software. Duration of the sounds was 1028 msec for 'kalisu' and 1029 msec for 'mufapi'. Both speech signals were converted into three-tone SWS stimuli (replacing F1, F2 and F3 by sine-waves) by a script from C. Darwin (http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/SWS) run in Praat, a speech analysis/synthesis software [31].

The videos showed the speaker's face against a dark background and were converted into bitmap sequences, which were matched on total duration (46 frames ~1535 msec) and auditory onset of the first syllable. There were two inter-stimulus differences in terms of timing: the onset of the second syllable in 'kalisu' (i.e., /li/) lagged the onset of /fa/ in 'mufapi' with 16 msec whereas the onset of the third syllable in 'kalisu' (/su/) was 229 msec earlier than the onset of /pi/ in 'mufapi'. These internal timing differences were introduced to serve as a temporal cue to the mismatch between the sound and the *incorrect* video, given that it was larger than the adult temporal window of integration [e.g., 10, 11, 32].

2.3. Procedure and design; Adults

Participants were seated in a sound-attenuated and dimly lit booth in front of a 19 -inch monitor. A keyboard was used for data acquisition and sounds were delivered through regular computer speakers centered beneath the screen. During a trial, the two videos were presented simultaneously, one on the left side, the other on the right, while a naturally timed sound (natural speech in the NS group and SWS in the SWS group) that matched one of the two videos was delivered. Counterbalancing of sound identity ('kalisu' or 'mufapi') and the side of the matching video (left or right) yielded 4 different conditions, all repeated 12 times (48 trials in total) in random order. After each trial, participants were asked to indicate whether the sound matched the left or right video by pressing a corresponding key. Importantly, the experimental instruction made no reference to the fact that SWS sounds were derived from speech.

2.4. Procedure and design; Infants

Infants sat on a caregiver's lap in a dimly lit testing booth at approximately 100 cm in front of two 19-inch computer monitors used for stimulus presentation. These monitors were placed 5 cm apart in a 170°-angle. Caregivers were instructed not to speak and to refrain from moving as much as possible during the experiment. The experiment was run from a laptop (Dell Latitude E4310) that controlled the two monitors. The videos were 17(H) x 14(W) cm in size and spacing between the centers of the left- versus right articulating mouths was 65 cm. A third monitor (placed behind the two screens that presented the stimuli) displayed an initial fixation stimulus and was controlled by a PC. Speech sounds were delivered through a regular PC speaker that was placed behind the screens and a second speaker delivered sounds during fixation. Infants' looking behavior was recorded by a digital video camera (Canon FS300) that was centered between the front screens (see Figure 1).

The experiment was ~2 min in duration and consisted of three phases: a fixation phase, a visual-only familiarization phase to acquaint them with the display, and an audiovisual preferential looking procedure. Sound identity ('kalisu' or 'mufapi'), location of visual familiarization start (left or right screen), speech type (natural speech or SWS), and location of the matching video during testing (left or right) were counterbalanced across participants.

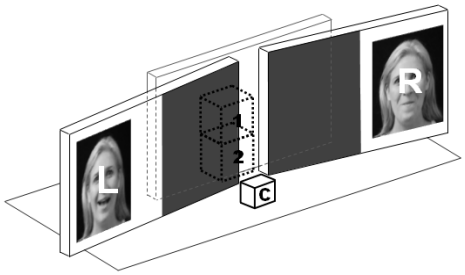

Figure 1: *Overview of the experimental set-up for infants. The left- and right screen presented stimuli and were placed in front of the middle screen that was used to direct gaze towards midline (fixation). Looking behavior was recorded with a camera (c) and speakers 1 and 2 presented sound during fixation and stimulus presentation respectively.*

2.5.1 Fixation

Color-alternating videos of geometrical shapes were presented in combination with an attractive sound (i.e. a squeeze-toy sound, a bicycle bell or a toy-car honk) until a live feed from the camera confirmed that infants' attention was directed towards midline.

2.5.2 Familiarization

Infants were familiarized with the dual-screen procedure by being exposed to one (silent) video ('kalisu' or 'mufapi') on either the left or the right screen a total of three times (ISI = 500 msec), while the other screen was black. Next, the other video was displayed on the opposite side following the same procedure. Finally, three repetitions of both videos were delivered simultaneously on both screens followed by a 1750 msec period in which both screens were black.

2.5.3 Preferential looking

Both videos were presented simultaneously 36 times (i.e. 36 trials, ITI = 500 ms) in the same locations as during familiarization, while a naturally-timed sound was played (natural speech or SWS) that matched one of the two videos.

3. Results

3.1. Adults

For each adult, the proportion of 'correct'-responses (i.e., the selected video corresponded with the sound) was averaged across all 48 trials. As can be seen in Figure 1, participants in the natural speech group reached ceiling after ~4 trials (mean proportion of correct responses was .96) while participants in the SWS group performed significantly worse (mean proportion

correct-responses was .71, t[50] = 6.07, $p < .001$) with little improvement in the second half of trials.

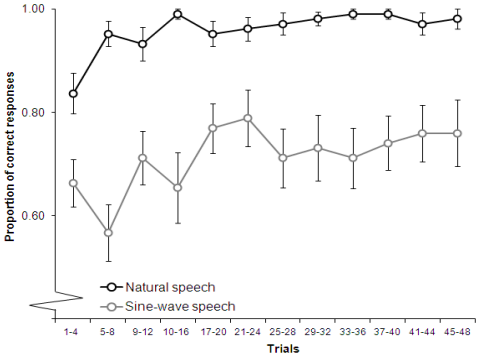

Figure 1: *The proportion of correct responses for adults in the natural speech and SWS groups averaged across bins of 4 trials. The error bars represent one standard error of the mean.*

These results indicate that adults were well able to match natural speech sounds to corresponding articulating faces and were better able to do so than adults who heard SWS.

3.2. Infants

The camera footage obtained during the experiment was stored for off-line analysis. Frame by frame inspection of all footage was done by two observers (MB and KS). Reliability of the coding was assessed by computing inter-observer Spearman's rank order correlations for the time spent looking at the matching screen, the non-matching screen and time spent not looking at the screens (all p-values < .001).

Infants' looking behavior was averaged across speech sound identity, starting-location of the visual familiarization phase, and location of the screen that matched the audio (Left vs. Right) by computing proportions of time spent looking at the screen that matched – and did not match – the audio, as well as the proportion of time infants did not look at the screens (i.e., 'Other'). These data are depicted in Table 1. A 2 (Gaze direction; Matching- versus Non-matching screen) * 2 (Speech type; Natural speech versus SWS) ANOVA that showed a main effect of Gaze direction as the proportion of looking at the screen that matched the audio was significantly higher than the proportion of time spent looking at the screen that did not match the audio ($F[1,16] = 12.97$, $p < .003$). There was no interaction between Gaze direction and Speech type ($F < 1$) and no main effect of Speech type ($F[1,16] = 2.07$, $p = .17$).

Speech type	Proportions of looking times			
	Matching	Non-matching	Other	PLM
Natural	.55	.26	.19	.68
SWS	.62	.27	.11	.69

Table 1: *Infants' proportions of looking times at the matching- and non-matching screen and the proportion of time infants were not looking at the Screens (Other). The final column displays the PLM values that were calculated by dividing the proportion 'match' by the total proportion of looking to the two screens.*

Next, we conducted PLM-values (Proportion of looking at the screen that matched the audio / Total proportion of looking at the screens) and compared these values against 50% chance-level. The PLM values of .68 and .69 for the infants that heard natural speech versus SWS, respectively, were significantly higher than chance (p-values < .043).

4. Discussion

Observers may use a combination of temporal and phonetic cues to integrate AV speech.

For adults, performance improved drastically when sounds were natural speech rather than SWS, most likely because the AV phonetic cues in natural speech signal were beneficial for correspondence detection.

In contrast, infants did not seem to benefit from phonetic cues when detecting the AV correspondence. Instead, they presumably only relied on the AV temporal correlation between the sound and the matching video, as the AV asynchrony between the sound and the non-matching video was likely to be too small to be detected [e.g., 9, 33].

It is demonstrated that, the infant auditory system is sensitive to 25 msec temporal modulations [34], which correspond to the temporal detail needed to extract segmental information from the speech signal [35]. More specifically, when comparing 12 msec modulations with 25 msec modulations, Telkemeyer and colleagues [34] reported enhanced brain activity for the latter in neonates' bilateral inferior and posterior temporal brain regions, as well as the in the right temporoparietal region, a brain area demonstrated to be sensitive to auditory sequences with temporal structure similar to speech syllables [36].

We therefore assume that infants were able to detect the temporal auditory structure in the stimuli in detail, and we propose that their AV matching was guided by the cross-modal temporal correlation. This inference seems plausible for 2 additional reasons: 1) infants cannot match a static artificial three-tone complex onto visual speech [37] whereas, as demonstrated here, infants can match three-tone complexes that share the temporal relationship that exists between natural speech with visual speech, and 2) the ability to extract phonetic information from the lipread signal appears increases over developmental time [6, 16-21]. For instance, the visual bias on sound identification in children is reported to be less than 10%, up to 57%, of the visual bias in adults [16-18, 38].

This does not imply that infants are not sensitive to phonetic information. In fact, it has been demonstrated that infants do integrate auditory and visual speech on a phonetic level [e.g., 4], although this process is not mandatory for infants [22] and may be distinct from mere correspondence detection. Given the above, we would like to propose that infants mainly rely on the temporal cues in the signal when detecting AV correspondence, whereas adults clearly benefitted from additional phonetic cues.

However, the infant data presented here is based on a rather small sample so any definitive conclusions are not yet in order. Moreover, a systematic degradation of the speech signal in time and phonetic detail is needed to chart the developmental trend underlying the relative contributions of these cues for speech perception in more detail.

5. References

[1] P. K. Kuhl and A. N. Meltzoff, "The bimodal perception of speech in infancy," *Science*, vol. 218, pp. 1138-1141, 1982.

[2] M. L. Patterson and J. F. Werker, "Matching phonetic information

in lips and voice is robust in 4.5-month-old infants," *Infant Behavior and Development*, vol. 22, pp. 237-247, 1999.
[3] L. D. Rosenblum, M. A. Schmuckler, and J. A. Johnson, "The McGurk effect in infants," *Perception & Psychophysics*, vol. 59, pp. 347-357, 1997.
[4] D. Burnham and B. Dodd, "Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect.," *Developmental Psychobiology*, vol. 45, pp. 204-20, Dec 2004.
[5] P. K. Kuhl and A. N. Meltzoff, "The intermodal representation of speech in infants," *Infant Behavior and Development*, vol. 7, pp. 361-381, 1984.
[6] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, 1976.
[7] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215, 1954.
[8] D. J. Lewkowicz, "Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables.," *Child Development*, vol. 71, pp. 1241-57, 2000 Sep-Oct 2000.
[9] D. J. Lewkowicz, "Infant perception of audio-visual speech synchrony.," *Developmental Psychology*, vol. 46, pp. 66-77, Jan 2010.
[10] A. Vatakis and C. Spence, "Audiovisual synchrony perception for music,speech,and object actions.," *Brain research*, vol. 1111, pp. 134-142, 2006.
[11] K. W. Grant, V. van Wassenhove, and D. Poeppel, "Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony," *Speech Communication*, vol. 44, pp. 43-53, 2004.
[12] P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech perception in infants," *Science*, vol. 171, pp. 303 - 306, 1971.
[13] E. Kushnerenko, T. Teinonen, A. Volein, and G. Csibra, "Electrophysiological evidence of illusory audiovisual speech percept in human infants.," *Proeedings of thevNationall Academy of Sciences of the United States of America,*, vol. 105, pp. 11442-5, Aug 2008.
[14] D. Burnham and B. Dodd, "Auditory-visual speech perception as a direct process: The McGurk effect in human infants and across languages.," in *Speechreading by humans and machines*, D. G. Stork and M. E. Hennecke, Eds., ed Berlin: Springer-Verlag, 1996, pp. 103-114.
[15] P. W. Jusczyk and P. A. Luce, "Infants′ Sensitivity to Phonotactic Patterns in the Native Language," *Journal of Memory and Language*, vol. 33, pp. 630–645, 1994.
[16] D. W. Massaro, "Children's perception of visual and auditory speech," *Child Development*, vol. 55, pp. 1777-1788, 1984.
[17] N. S. Hockley and L. A. Polka, "A developmental study of audiovisual speech perception using the McGurk paradigm.," *Journal of the Acoustical Society of America*, vol. 96, p. 3309, 1994.
[18] R. N. Desjardins, J. Rogers, and J. F. Werker, "An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks," *Journal of Experimental Child Psychology*, vol. 66, pp. 85-110, 1997.
[19] K. Sekiyama and D. Burnham, "Issues in the development of auditory-visual speech perception: Adults, infants, and children.," ed. Paper presented at the 8th International Conference on Spoken Language Processing, Jeiu Island, Korea., 2004.
[20] V. Bruce, R. N. Campbell, G. Doherty-Sneddon, A. Import, S. Langton, S. McAuley, *et al.*, "Testing face processing skills in children," *British Journal of Development Psychology*, vol. 18, pp. 319–333, 2000.
[21] L. A. Ross, S. Molholm, D. Blanco, M. Gomez-Ramirez, D. Saint-Amour, and J. J. Foxe, "The development of multisensory speech perception continues into the late childhood years," *European Journal of Neuroscience*, vol. 33, pp. 2329-37, Jun 2011.
[22] R. N. Desjardins and J. F. Werker, "Is the integration of heard and seen speech mandatory for infants?," *Developmental Psychobiology*, vol. 45, pp. 187-203, 2004.
[23] J. L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, pp. B69-78, Sep 2004.
[24] K. Eskelund, J. Tuomainen, and T. S. Andersen, "Multistage audiovisual integration of speech: dissociating identification and detection," *Experimental Brain Research*, vol. 208, pp. 447-457, Feb 2011.
[25] J. Vroomen and J. J. Stekelenburg, "Perception of intersensory synchrony in audiovisual speech: Not that special," *Cognition*, vol. 118, pp. 75-83, Jan 2011.
[26] R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, pp. 947-9, May 22 1981.
[27] J. Tuomainen, T. S. Andersen, K. Tiippana, and M. Sams, "Audio-visual speech perception is special," *Cognition*, vol. 96, pp. B13-22, May 2005.
[28] J. Vroomen and M. Baart, "Phonetic recalibration only occurs in speech mode," *Cognition*, vol. 110, pp. 254-9, Feb 2009.
[29] M. L. Patterson and J. F. Werker, "Two-month-old infants match phonetic information in lips and voice," *Developmental Science*, vol. 6, pp. 191-196, 2003.
[30] G. Hollich, R. S. Newman, and P. W. Jusczyk, "Infants' use of synchronized visual information to separate streams of speech.," *Child Development*, vol. 76, pp. 598-613, 2005 May-Jun 2005.
[31] P. Boersma and K. Weenink, "Praat: doing phonetics by computer, Retrieved from http://www.fon.hum.uva.nl/praat," ed, 2005.
[32] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Temporal window of integration in auditory-visual speech perception," *Neuropsychologia*, vol. 45, pp. 598-607, 2007 2007.
[33] D. J. Lewkowicz, "Perception of auditory-visual temporal synchrony in human infants.," *Journal of Experimental Psychology; Human Perception and Performance*, vol. 22, pp. 1094-106, Oct 1996.
[34] S. Telkemeyer, S. Rossi, S. P. Koch, T. Nierhaus, J. Steinbrink, D. Poeppel, *et al.*, "Sensitivity of newborn auditory cortex to the temporal structure of sounds.," *The Journal of Neuroscience*, vol. 29, pp. 14726-33, Nov 2009.
[35] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects.," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 336, pp. 367-73, Jun 1992.
[36] F. Homae, H. Watanabe, T. Nakano, and G. Taga, "Functional development in the infant brain for auditory pitch processing.," *Human Brain Mapping*, vol. 33, pp. 596-608, Mar 2012.
[37] P. K. Kuhl, K. A. Williams, and A. N. Meltzoff, "Cross-modal speech perception in adults and infants using nonspeech auditory stimuli," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 17, pp. 829-840, 1991.
[38] D. W. Massaro, L. A. Thompson, B. Barron, and E. Laren, "Developmental changes in visual and auditory contributions to speech perception.," *Journal of Experimental Child Psychology*, vol. 41, 1986.