Baby steps in perceiving articulatory foundations of phonological contrasts: Infants detect audio→video congruency in native and nonnative consonants

Catherine T. Best[1,2,3], Christian H. Kroos[1], Sophie Gates[1] and Julia Irwin[2,4]

[1]MARCS Institute, U Western Sydney, Australia [2]School of Humanities & Communication Arts, U Western Sydney, Australia [3]Haskins Laboratories, USA [4]Southern Connecticut State U, USA

c.best@uws.edu.au chkroos@gmail.com s.gates@uws.edu.au irwinj1@southernct.edu

Theoretical background. Whether the perceptual primitives that perceivers use to recognize phonological distinctions include articulatory information or, as widely assumed, only acoustic information remains an important topic of debate. The answer is central to understanding infants' early steps toward acquisition of their native phonology. Compatible with the premise that perceivers detect articulatory information in speech, adults' perception of acoustic speech is systematically influenced by visible speech movements: a synchronously-presented phonetically-congruent talking face enhances intelligibility of speech in noise, whereas a phonetically-incongruent face can produce "misperception" of the audio token's place of articulation (McGurk effect). Similar results have been found in young infants. But bimodal audio-visual (AV) studies alone cannot pinpoint whether perceivers detect amodal *articulatory correspondences* between audio and video signals, or instead rely on *learned AV associations* between specific audio speech categories and the co-occurring face motions they have experienced with those sounds. Resolving this issue requires examination of: 1) *intermodal* recognition (cross-modal "transfer") of the common information between audio-only (A) and video-only (V) tokens of a given speech target; and of 2) developmental changes in recognition of experienced (native) versus non-experienced (nonnative) speech contrasts. The only prior such study tested English vs. Spanish infants' detection of A→V congruence in a visible contrast used in English but not Spanish: /b/-/v/. Both groups succeeded at 4 months but only English infants did so at 11 months (Pons, Lewkowicz, Sebastián-Gallés & Soto-Faraco, 2009). The 4-month results refute the hypothesis that perception of intermodal congruence rests on learned A-V associations. However, because both stimuli were labials (same articulator: LIPS) we can't tell whether they detected AV congruence based on articulator-distinct or other information (e.g., phonetic features; cross-modal psychophysical properties). Moreover, as Spanish 11-month-olds fail to discriminate audio-only /b/-/v/, their lack of differential response to the silent /b/ vs. /v/ videos after being habituated to audio /b/ or /v/ tells us nothing about their ability to detect intermodal congruencies. Therefore, we designed a series of studies on 4- and 11-month-old infants' detection of intermodal A→V congruency for visibly-distinct between-articulator contrasts (LIPS-TONGUE TIP) in native and two types of nonnative consonants.

Experiments. Three experiments assessed sensitivity to intermodal A→V congruence in Australian English-learning infants of 4 months (n = 20/expt) and 11 months (n = 20/expt). The audio-only stimuli used for the habituation phase of the task were natural tokens produced by native female speakers of each target language: English stops /pa/-/ta/ (Expt 1); Tigrinya ejectives /p'a/-/t'a/ (Expt 2); !Xóõ bilabial vs. dental clicks /ʘa/-/ǀa/ (Expt 3). The Tigrinya contrast was chosen because it has been shown that English adults and infants discriminate the audio contrast easily, but that adults nonetheless clearly hear these items as non-English-like (e.g., "choked" or "coughed" deviant variants of P vs. T). The Xóõ clicks were chosen for comparison, as prior findings indicate English adults hear them as nonspeech articulations ("kiss" vs. "tsk" sounds). In a separate study with Australian 4- (n = 30) and 11-month-olds (n = 25) we confirmed that both ages can discriminate audio-only tokens of these clicks. The silent-video stimuli in the A→V experiments were of an American female producing English /ba/ and /da/, which match the place of articulation difference for each target contrast, yet use a different speaker and voicing feature all three audio contrasts. The A→V conditioned visual fixation task had four phases: 1) habituation to the bilabial or coronal audio item (with on-screen checkerboard display; 2 subgroups/age of n = 10); 2) two "lag" trials of additional audio target presentations to assure habituation was maintained; 3) two video-only "buffer" trials of the video talker blinking her eyes, to familiarize infants to the switch to silent video stimuli; 4) test phase in which two trials presented the silent-video bilabial vs. coronal articulations. Intermodal sensitivity was evaluated by comparing infants' fixation times to the video trials that were articulator-congruent vs. articulator–incongruent to the infant's audio habituation stimulus, via ANOVAs at each age/experiment.

Results and Discussion. Four-month-olds showed a fixation preference for A→V articulator congruency in both the English and Tigrinya contrasts, but 11-month-olds only preferred A→V congruency for English, reversing to a novelty preference, i.e., for A→V *in*congruency, with the Tigrinya ejectives. In striking contrast, the 4-month-olds showed an A→V *in*congruency preference for the !Xóõ clicks, while the 11-month-olds failed to detect any A→V congruency relationship. We propose that both 4- and 11-month-olds recognize the Tigrinya ejectives as speech articulations, but only the older group recognize that they are deviant from English. In contrast, 4-month-olds already recognize !Xóõ clicks as deviant from English, while 11-month-olds instead treat them as nonspeech, like adults. We will discuss the implications of these findings for theories of native-language perceptual attunement and phonological development.