FIRST LANGUAGE
VOLUME 28 - ISSUE 1 - NUMBER 84 - FEBRUARY 2008

Developmental and cognitive aspects of children's disbelief comprehension through intonation and facial gesture

Journal:	*First Language*
Manuscript ID	FLA-17-0035.R2
Manuscript Type:	Original Manuscript
Keywords:	prosody, intonation, belief states, facial gesture, Theory of mind, belief reasoning, false belief
Abstract:	We investigate how children leverage intonational and gestural cues to an individual's belief state through unimodal (intonation-only or facial gesture-only) and multimodal (intonation + facial gesture) cues. A total of 187 preschoolers (ages 3-5) participated in a disbelief comprehension task and were assessed for Theory of Mind (ToM) ability using a false belief task. Significant predictors included Age, Condition and success on the ToM task. Performance improved with age, and was significantly better for the multimodal condition compared to both unimodal conditions, suggesting that even though unimodal cues were useful to children, the presence of reinforcing information for the multimodal condition was more effective for detecting disbelief. However, results also point to the development of intonational and gestural comprehension in tandem. Children that passed the ToM task significantly outperformed those that failed it for all conditions, showing that children who can attribute a false belief to another individual may more readily access these intonational and gestural cues.

Developmental and cognitive aspects of children's disbelief comprehension through intonation and facial gesture

Abstract: We investigate how children leverage intonational and gestural cues to an individual's belief state through unimodal (intonation-only or facial gesture-only) and multimodal (intonation + facial gesture) cues. A total of 187 preschoolers (ages 3-5) participated in a disbelief comprehension task and were assessed for Theory of Mind (ToM) ability using a false belief task. Significant predictors included Age, Condition and success on the ToM task. Performance improved with age, and was significantly better for the multimodal condition compared to both unimodal conditions, suggesting that even though unimodal cues were useful to children, the presence of reinforcing information for the multimodal condition was more effective for detecting disbelief. However, results also point to the development of intonational and gestural comprehension in tandem. Children that passed the ToM task significantly outperformed those that failed it for all conditions, showing that children who can attribute a false belief to another individual may more readily access these intonational and gestural cues.

Introduction

Quite a few commonalities can be found when we consider the role of speech prosody and facial gesture in a child's early development. Here we take the term prosody to refer to the continuous changes in speech that affect duration, intensity and pitch patterns (for an overview see Ladd 2008). Kendon (2004) refers to facial gestures as "eyebrow movements or positionings, movements of the mouth, head postures and sustainments and changes in gaze direction" (p. 310)[1]. In conversation, as Bavelas, Gerwing and Healing (2014) point out, these gestural movements are synchronized with speech, both in timing and meaning. Babies respond to both prosody and facial gesture from early on. In fact, human neonates have access to prosodic information *in utero*, with recent studies showing that even infant cries reflect the prosodic patterns of a child's ambient language (Mampe, Friederici, Christophe & Wermke, 2009; Wermke et al., 2016). Newborns have been shown to use prosodic information to discriminate between languages from different rhythmic classes (Nazzi, Bertoncini & Mehler, 1998). Cooper and Aslin (1990) found that both newborns and 1-month-old infants have a preference for infant-directed speech, a speech style that includes prosodic modifications that include higher overall pitch, slower tempo, longer pauses and increased focus-marking. Unlike prosody, however, information about human faces is not available to human neonates during the gestation period. Even so, information on the face becomes important to infants very early on. For instance, at 2 months of age, infants are able to discriminate happy vs. neutral faces in holographic stereograms (Nelson & Horowitz, 1983) and 3-month-olds are able to discriminate happy, sad and surprised faces (Young, Brown, Rosenfeld & Horowitz,

[1] In the literature that explores the relationship between cues on the face and emotion, the term *facial expressions* is typically used, while the literature exploring cues on the face as related to speech use the term *facial gestures.* Since the types of facial cues we explored in this study are expected to be synchronized with speech, we also employ the term facial gesture to refer to our object of study.

1977) as well as smiling vs. frowning faces (Barrera & Maurer, 1981) from photographic stimuli.

Early access to meaning through prosody and facial gesture

In terms of infants' ability to extract *meaning* from prosody or facial gesture, most of the literature has been focused on affective or emotional meaning. Five-month-olds demonstrate positive affect when hearing approving utterances vs. prohibiting statements that differ in F0 patterns. They have been shown to demonstrate negative affect when they heard target prohibitive statements in both their L1 as well as unfamiliar languages (Fernald, 1993). Using event-related potentials, Grossman, Striano and Friederici (2005) showed that 7-month-olds allocated more attention to angry prosody, showing evidence that infants differentiate their attentional responses based on the prosodically-conveyed emotional valence present in the stimuli. The 12-month-olds in Mumme, Fernald and Herrera's (1996) study also showed negative affective behavior in response to negative affect prosody in exclamatives with the word *Oh!* in English. Thus within the first year of life, babies have at least some access to prosodic meaning, albeit rudimentary.

As noted above, while prosodic information is available to fetuses during the gestational period, cues from the face are not. Prosodic cues and facial cues are certainly not integrated during the gestational period. Despite these facts, the findings for infants' early access to the meanings of facial gesture are similar to those related to prosody. By six months, infants may react more negatively (e.g. frowning or crying) to sad and angry facial expressions when compared to happy or neutral facial expressions (Kreutzer & Charlesworth, 1973). Gaze patterns are of notable importance, since infants engage in joint attention patterns and use gaze alternation to confirm that a person is attending to a target (e.g., Trevarthen & Hubley, 1978; Mundy & Newell, 2007). Scorce, Emde, Campos and Klinnert (1985) found that 12-month-olds are more likely to approach a visual cliff when their mothers showed a happy face, but retreated when their mothers produced fearful faces. Thus, infants seek out cues from the face and may base their own behavior on these cues (Klinnert, 1984). In spite of this early ability to imitate, discriminate and react to information on the face, Nelson (1987) suggests that between approximately one and two years of age, young children's knowledge of facial gesture is quite rudimentary in that they are familiar with only basic emotions, and often only a subset of these. Based on studies of facial gestures in primates, he suggests that at least some component of the ability to recognize cues from the face could be innate, but that this ability would then be modified by experience. Nevertheless, Nelson claims that the ability to understand facial gestures "undergoes a long incubation period in the human" (p. 906).

While the present study focuses on how older children (ages 3-5) interpret meanings conveyed through prosody (specifically intonation) and facial gesture, it appears that infants gain access to some types of meaning, specifically emotional meaning, at very young ages. However, there is some evidence that when a child starts forming a lexicon, the way prosody and facial gestures are used in comprehension becomes affected. Friend (2001) explored 15- and 16-month-olds' sensitivity to prosody as well as facial gestures versus lexical content. In this task, children who were about to play with a novel object saw videos of a speaker with either an approving or disapproving message. Friend found that receptive vocabulary was a significant predictor of children's behavior: the children who understood the lexical meaning of the message were better regulated by lexical content than by prosody or facial gesture. On the other hand, younger children were better regulated by prosody and facial gesture. This finding is also consistent with work from Lawrence and Fernald (1993), who showed that 9-month-olds were better regulated by tone of voice compared to lexical content while the reverse was true for 18-month-olds. Friend proposes a transition stage from affective to linguistic meaning around the age of 15 months. Thus, as children get older, the extent to which they rely on prosody and/or facial gestures to guide them to specific meanings may change. Additionally, it is not clear how older children gain access to meaning associated with these modalities outside the realm of emotions.

Prosody and facial gestures are thus sources of information that babies pay attention to in early developmental stages. From those sources, babies can access types of information about individuals' emotions. The parallels between prosody and facial gesture, however, seem to change as children get older. Comprehension studies have shown a clear advantage for gestures (including facial gestures) over what has been referred to in many studies as 'vocal cues'. For instance, Nelson and Russell (2011) carried out an experiment where preschoolers (ages 3-5) had to label emotions (happiness, sadness, anger and fear) based on video clips produced with four different cue conditions: face-only, body posture-only, voice-only and multi-cue (i.e. face + body + voice). Results showed that most children did not choose the correct label for the stimulus presented for the voice-only condition. However, labels for the face-only condition did not differ significantly from the multi-cue condition and labels for the multi-cue condition were significantly more accurate when compared to the body posture condition. However, recent work by Nelson and Russell (2016) showed that children may often use the process of elimination in labeling tasks, and warn that previous studies may overestimate children's facial expression knowledge, and children's apparent recognition of emotion from facial gesture may be an 'artifact of method' (p. 62).

Intonation, facial gesture and belief states

Here we use the term *gesture* as a broad term referring to the use of the hands or other parts of the body for communication. Thus *facial gesture* would be a subtype of this term. While the role of gesture for language acquisition is well studied for hand gestures, (McNeil, Alibali & Evans, 2000; Demir, Fischer, Goldin-Meadow & Levine 2014), less is known about the role of exploring facial gesture in a child's linguistic development, which includes a child's intonational development. As we have pointed out, the bulk of the work on early access to prosodic meaning and facial gesture meaning is related to individuals' *emotions*. On the other hand, the work focusing on the facilitating role of gesture in comprehension has been related to lexical comprehension or the comprehension of complex syntactic messages, rather than intonational meaning. In the present paper we were interested in how children might use intonation, as well as facial gesture to calculate speaker *belief states.* Emotional states and belief states are similar in that they are *internal* states of the speaker, but the latter deals with *epistemic* aspects of language such as degree of certainty or uncertainty about propositional content. Specifically, we focused on children's comprehension of an individual's state of **disbelief**. We asked to what extent children are able to infer an individual's state of disbelief through different modalities: prosody (specifically intonation), facial gesture, or the combination of the two. Similar to what has been found for the case of prosody and emotion, the preschool and early school years have also been shown to be an important developmental window for children's ability to comprehend intonational forms associated with speaker belief states. Armstrong (2014) investigated children's comprehension of prosodically-encoded *disbelief,* i.e. when a speaker expresses her inability to believe some proposition, as in (1):

(1)
A: I just fed the rhinoceros in the living room. He's so cute!
B: There's a rhinoceros in the living room?!?

Rhinoceroses are not typically pets and certainly not known to frequent people's living rooms in Western culture. Thus, B expresses her state of disbelief, or inability to accept the (p)roposition *There is a rhinoceros in the living room* into her set of beliefs. Many languages mark this disbelief meaning (conveyed orthographically in (1) as ?!?) with an intonational morpheme. Armstrong (2014) looked specifically at how child speakers of Puerto Rican Spanish were able to comprehend disbelief meaning as conveyed by the L* HL%[2] contour, the intonational morpheme for marking disbelief in questions in Puerto Rican Spanish. In order to test this, a task was designed using a Powerpoint presentation featuring a set of twins and their friend, Jeni. Jeni was telling the twins about the animals she saw while she was on vacation. The child was told that there was always one twin who did not believe that Jeni saw the animal she claimed to have seen on vacation, and that they would know which twin that was by listening carefully to what the twins said. Thus when Jeni said *Yo vi un búho.* 'I saw an owl', the child heard one twin reply *¿Un búho?* 'An owl?' with neutral echo question intonation, produced with ¡H* L%[3], while the other twin asked the same question, but with disbelief intonation, produced with L* HL%. Results showed that while 4- and 5-year-olds performed at above-chance levels on the task, 6-year-olds significantly outperformed both groups. Interestingly, some 6-year-olds produced facial gestures known to be associated with polar questions when they heard stimuli produced with ¡H* L%, and facial gestures known to be associated with disbelief when they heard stimuli produced with L* HL%. This suggests that children may strongly associate specific facial gestures with certain intonational melodies. As mentioned above, Hübscher, Esteve-Gibert, Igualada and Prieto (2017) also investigated children's comprehension of intonation related to a speaker's belief state, more specifically, their degree of certainty. Using the same procedure described above, the authors found that 3-5-year-old Catalan-speaking children are better at comprehending uncertainty when some sort of facial gesture cue is present. However, they also found that 3-year-old children were more sensitive to intonational cues to uncertainty compared to lexical cues (such as *maybe*). This shows that by 3 years of age children have learned something about the relationship between prosody and belief states. These authors also found that both younger and older children performed better in detecting uncertainty when visual cues (e.g. facial gestures related to uncertainty) were present, and suggest

---

[2] This is transcribed using the Sp_ToBI system, the prosodic transcription system for Spanish. See Hualde & Prieto (2015) for the most recent Sp_ToBI labeling conventions.
[3] Also transcribed using the Sp_ToBI labeling conventions.

that visual information may help bootstrap children into linguistic meaning, as has been proposed in other work (Kelly, 2001; Butcher & Goldin-Meadow, 2000; McNeill, Cassel & McCullough, 1994). In terms of production, Armstrong (2018) showed that by the second half of the third year of life, two Puerto Rican Spanish-acquiring toddlers had produced some type of belief marking intonation within the question domain, though it is unclear to what extent these types of questions are comprehended at that age.

In earlier work, Moore, Harris and Patriquin (1993) compared the ability of children aged 3 to 6 to comprehend degrees of certainty conveyed through prosody vs. mental state verbs like *think*, *guess* and *know*. The youngest children could not use either type of cue, while older children showed an advantage for lexical information over prosody. However, the authors stress the fact that children in this age group are developing the ability to make inferences about mental states as conveyed through prosody and the lexicon. They suggest that in order to do so, a child's 'representational Theory of Mind' must be developed to a certain degree in order to comprehend mental state language, regardless of whether it is expressed prosodically or lexically.

Not unlike prosodic comprehension, facial gesture comprehension also continues to develop during the preschool and early school years. Nelson, Widen and Russell (2007) found that children are beginning to be able to identify a surprised face, which is a belief-related state, during the preschool years. Thus even though visual information may aid children in the detection of linguistic meaning, this does not mean that their ability to use information from the face is completely adult-like. Widen & Russell (2008), argue that children are "fine-tuning" their way of interpreting faces between these ages, based on labeling studies for children between the ages of 2 and 5.

Predictors in children's ability to comprehend belief states

Theory of Mind (ToM) refers to an individual's cognitive ability to attribute mental states to themselves and to other individuals. Such attributions may be verbal or non-verbal (Goldman, 2012). Children learn to become adept at using different sources of information, be it linguistic or extra-linguistic, as evidence for the mental states of others. One common way of assessing a child's developing ToM is the false belief task (Wimmer & Perner, 1983), which measures a child's ability to perceive that other individuals have beliefs that differ from each other. Success on such a task has been shown to be related to children's acquisition of the language associated with belief states. Since we saw belief reasoning as quite important for our comprehension task, the study described below includes a variation of the false belief task carried out by Wimmer and Perner.

Goals and research questions

As we have mentioned, the goal of our study was to understand how children access a speaker's state of disbelief through intonation, facial gesture, or the combination of these two cues. Prior work has suggested a bootstrapping effect for facial gesture, meaning that in the acquisition process children may first acquire facial gesture meaning, which may in turn give them access to intonational meaning. We thus hypothesized that children would perform better when detecting disbelief based on facial gesture when compared to intonation. In this case, we would also expect better performance when both cues are present. However, we also expected to find improvement on our task with age. Since disbelief is better perceived through facial gesture compared to intonation this difference is likely to diminish with age. Further, we hypothesized that the more sophisticated a child's belief reasoning skills (i.e. more developed representational Theory of Mind), the better they would be at detecting an individual's state of disbelief. Our experimental design for testing these hypotheses is detailed below.

Methods

Participants

A total of 187 Central Catalan[4]-speaking children (89 female and 98 male), mean age 4;5 (ranging from 2;10 to 6;3) participated in the study. Thirty Central-Catalan speaking adults participated as controls in this study. The total sample of 187 participants consisted of three grade levels, based on the structure of the Catalan school system: grades P3, P4 and P5, which are largely linked to a child's age. The child participants were recruited from schools[5] in Catalonia within a 1-hour radius of Barcelona. The children's parents filled out a language background questionnaire and signed a consent form. Parents were asked to report what percentage of the day their child spent communicating in Catalan. In order to be included in the study, a minimum of 80% Catalan usage had to be reported for a child to be included. Parents reported no language or hearing disorders for the participants.

Materials

Disbelief comprehension task
Three types of stimuli were prepared for the disbelief comprehension task: audio only stimuli (AO), visual only stimuli (VO) and combined audiovisual stimuli (AV). The AO stimuli were extracted from videos using Quicktime, and the audio portion of the stimuli was removed using Adobe Premiere in order to prepare the VO stimuli. In terms of the image for the VO stimuli, still images of actors were extracted from the original AV stimuli.

To create the stimuli for the comprehension task, we videorecorded two native speakers of Central Catalan. To make the stimuli more realistic and relatable for the children, we recorded two child actors (a male and a female) for the comprehension task. The male was 13 years of age at the time of recording, and the female was 11 years of age. In order to best target the intonational contrast of interest, we used very short utterances during which the nuclear configurations and facial gestures were realized. All utterances were fragments, consisting of NPs with the structure Determiner + Noun and had indefinite articles in the determiner slot, for example *Una balena?* 'A whale?' Importantly, none of

[4] Catalan is a Romance language spoken in northeastern Spain: in Catalonia, the Valencian Community and the Balearic Islands. It is the official language of Andorra, and is also spoken in parts of France and Italy. There are two dialectal blocks of Catalan: Western and Eastern. Central Catalan is one of the four dialects pertaining to the Eastern block: Northern Catalan, Central Catalan, Balearic and Algherese (Prieto & Rigau, 2007). Central Catalan is the dialect spoken in the capital and largest city in Catalonia, Barcelona.

the lexical content of the target questions included any information about the meanings of interest. All target utterances were echo questions. To obtain the AV stimuli for the task, the child actors were given the question that needed to be recorded. They were asked to imagine they were in one of two situations. In the first situation (disbelief), the actors were asked to produce disbelieving echo questions with the L*LH% contour; in the second situation (asking for confirmation), the actors were asked to produce neutral echo questions with the nuclear configuration L+¡H* L% (following the Cat_ToBI description, see Prieto, 2014). Figures 1 and 2 show spectrograms and waveforms for the respective contours. Thus each echo question in the test items was recorded with each intonation contour of interest (L* LH% for disbelieving and L+¡H* L% for neutral) and both facial gestures of interest (disbelieving and general question-marking).

--INSERT FIGURE 1 ABOUT HERE--

--INSERT FIGURE 2 ABOUT HERE--

For the facial gestures, we asked the actors to produce brow furrowing, eyelid closure and forward movement of the head for the disbelief echo condition, and brow raising with eyes wide open for the neutral echo question condition (following adult patterns found in Crespo-Sendra, Kaland, Swerts & Prieto (2013)). Figure 3 shows representative still pictures of the facial gesture from the video clips used in the experiment as the two child actors uttered a neutral echo question (left panels) versus a disbelieving echo question (right panels).

--INSERT FIGURE 3 ABOUT HERE--

For all three conditions (AO, VO and AV), participants were presented with a Powerpoint presentation, shown in Figure 4. The Powerpoint always featured the set of twins on the lefthand side of the slide. The "twins" were created by duplicating either stills or videos of the same child actor, depending on the condition. Thus for the one female actor, a pair of twins was created (Emma and Aina), and for the one male actor, a pair of twins was created (Pau and Josep). On the righthand side of the Powerpoint, the twins' friend appeared. The friend was female for the female twins (named Laia), and male for the male twins (named Daniel). The premise of the scenario (following Armstrong, 2014) was that the twins' friend had just gone on vacation, and was telling them about the animals that they saw while they were on vacation. The stimuli were counterbalanced for order of presentation (whether the neutral or disbelieving question was presented first or second) as well as which twin produced which contour. We also counterbalanced based on whether the twin appeared on top or on the bottom of the Powerpoint slide.

--INSERT FIGURE 4 ABOUT HERE--

Our false belief task was a modified version of the Sally Ann task (Baron-Cohen, Leslie & Frith, 1985) and was presented in video form featuring two puppets[6]. Stills from the task are shown in Appendix A. In the video, a princess puppet appears in a scene where there were two covered containers. The princess states that she was hiding her ball where no one could find it, and puts the ball in the container on the right, covering it, as shown in (1a). The princess then announces that she is going to school, and leaves the scene. While the princess is gone, a lion puppet appears laughing in a mischievous way. He opens the container with the ball and observes that there is a ball in it. He looks in the other container and observes that there is nothing in it. He then takes the ball from the right container and puts it in the left container, covering it and saying "*Let's shut it"* (1b). After moving the ball and shutting both containers, the lion laughs again in a mischievous way and leaves. Finally, the princess returns, greeting the viewer, saying she is back from school (1c). Once the princess returns, the child was asked two questions (1) *On buscarà la pilota primer, la nena?* 'Where will the girl look for the ball first?' and (2) *On és la pilota, en realitat* 'Where is the ball, really?'. Children were given credit for making reference to the container on the right for question (1), and for making reference to the container on the left for question (2).

Procedure

Children were distributed across conditions in a between-subjects design. Sixty children received the AO condition (20 3-year-olds, mean age=3;4; 21 four-year-olds, mean age=4;3; 19 five-year-olds, mean age = 5;6). We obtained a total of 720 responses for this condition (6 test trials x 2 blocks x 60 participants = 720). Sixty-five children participated in the VO condition (23 3-year-olds, mean age=3;5; 22 4-year-olds, mean age=4;2; 20 five-year-olds, mean age=5;6). A total of 780 responses were obtained for this condition. Sixty-two children participated in the AV condition (21 3-year-olds, mean age=3;5; 21 4-year-olds, mean age=4;7; 20 five-year-olds, mean age=5;5). A total of 744 trials were obtained for this condition. Thus the total number of trials obtained, including all three conditions was 2244.

A control group of adults (10 Catalan-dominant adults per condition) did the experiment

[6] Full script is available at: http://blogs.umass.edu/armstrong/materials-2/

using an online survey format. Adult participants read the instructions themselves and performed the comprehension task in their homes. We confirmed that the adults had no problems identifying the correct answers to the test questions; they provided the correct answer 95.8% of the time for the AO condition, 92.5% of the time for the VO condition and 94.5% of the time for the AV condition.

In the experimental setting, the child was seated in front of the laptop computer, with the experimenter next to them. The experimenter had a score sheet where the participants' responses were annotated. The child was introduced to the set of twins on the Powerpoint slide (as in Figure 4), and was told that they were twins. They were then told that the twins had a friend named Laia (or Daniel, depending on the block the child received first). Laia (or Daniel) had just returned from vacation with his/her family and was telling the twins about the animals s/he saw. The child was then told that there was always one twin that did not believe the friend, and the child needed to identify which twin that was by closely listening to/looking at (depending on the condition) what each twin said. For example,

Experimenter:
La Marta els explica que va veure un mico. (monkey appears)
Llavors l'Emma li diu (plays soundfile 1)
I l'Aina li diu (plays soundfile 2)
Test question: *Quina bessona no es creu la Laia, la de dalt o la de baix? Assenyala-la.*

English translation:
Laia tells them that she saw a monkey. (monkey appears)
So Emma says to her (plays soundfile 1)
And Aina says to her (plays soundfile 2)
Test question: Which twin doesn't believe Laia, the one on top or the one on the bottom? Point to her.

After the test question, the child was asked to point to the twin they thought did not believe Laia. In instances where the child said "neither" the child was reminded that there was *always* one twin that did not believe the friend, and that they should do their best to decide which one it was. The child could listen as many times as they needed in order to make a decision. Each participant received two blocks of stimuli (Block 1 and Block 2), and one of two lists. For the first list, the child received all stimuli produced by the female actor in Block 1 and those produced by the male actor in Block 2. For list 2, participants received all stimuli from the male actor in Block 1, and the female actor in Block 2. Participants received four familiarization trials prior to Block 1, and two additional familiarization trials prior to Block 2 to familiarize them with the second speaker. For the familiarization trials, the same neutral versus disbelieving meanings were maintained, but the information was conveyed lexically rather than intonationally (and gesturally for the cases of VO and AV). Thus for the neutral condition, participants heard *Ah, què bé que veiessis una balena* 'Oh, that's good that you saw a whale', for a neutral reaction or *No m'ho crec, que veiessis una balena* 'I don't believe it, that you saw a dog'. There were six test trials per block, yielding a total of 12 test trials per child.

Results

ToM task

For the false belief task, 21% of children from grade P3, 73% from grade P4 and 88% of children from grade P5 passed the task. These results confirm findings from prior studies that between 3 and 4 years of age children improve significantly in their ability to pass a false belief task.

Disbelief comprehension task

Mixed model logistic regression

We fit a mixed logistic regression model for our data using the lmerTest packing in R (R Core Team, 2013) with Correct as the dependent variable (Correct vs. Incorrect) and with Age in months, Condition and Theory of Mind as fixed effects, as well as their interactions. Both Participant and Item were included as random factors[7]. Nested models were compared with the anova () function in R, and it was determined that the best fit model included all predictors but no interactions. Table 1 (A-C) shows the Estimates, Standard Error, z values and p values for our fixed effects, along with releveled versions of the model. Cells shaded grey indicate significant effects.

---INSERT TABLE 1 ABOUT HERE---

Tables 1A, B and C all show Age in months as a significant predictor. This is confirmed by the regression lines in Figure 5. Regardless of the baseline, ToM was always a significant predictor indicating that for each of the three conditions, participants who passed the ToM task performed significantly better (indicated by the higher positive Estimate) on the disbelief comprehension task. Table 1A shows that when performance on the AO task is compared to performance on the VO task, no significant difference is found, while performance on the AV task was significantly better compared to the AO condition. Table 1B, with the VO condition set as the baseline shows, again, the lack of significance when compared to the AO condition, but a significant result when the VO condition is compared to the AV condition. We can therefore conclude that performance on the multimodal condition (AV) was significantly better when compared to performance on either of the unimodal conditions (AO or VO). Again, as noted above, the best fit model did not include any significant interactions for our data.

----INSERT FIGURE 5 ABOUT HERE---

Discussion & Conclusions

Our results show that with age, the ability to perceive disbelief meaning, for all three conditions improves. Across ages, performance on the intonation + facial gesture (AV) condition was better when compared to the intonation-only (AO) or facial gesture-only

---

[7] The structure Correct ~ Age_months + ToM + Condition + (1 | Participant) + (1 | Item) was used for the best fit model.

(VO) conditions. This effect was confirmed in our statistical analysis, which showed that children performed significantly better on the intonation + facial gesture condition when compared to the intonation-only and the facial gesture-only conditions. While Figure 5 does show a trend that performance starts off better for the facial gesture-only condition when compared to the intonation-only condition, we did not find an interaction for this in our model nor did we find significant differences when comparing the two unimodal conditions. Our results show that children are continuing to develop the ability to perceive disbelief, through both intonation and facial gesture during the window of time we investigated, but that having both modalities present is beneficial to them. However, it should be noted that even when both cues are present, children continue to get better at the task with age. These findings support the idea that children are "fine-tuning" their ability to use facial gesture, and that this idea can be extended to intonation as well. Contrary to our hypothesis, however, when we compare the unimodal conditions (intonation-only vs. facial gesture only), we do not find that children performed significantly better on the facial gesture only condition. Instead, children were found to perform just as well on the intonation only condition. Our results suggest that during much of this important window (ages 3-5) intonation can be just as strong a cue to disbelief as facial gesture, and that these different cues to disbelief develop in tandem during this window.

As we note earlier in the paper, Hübscher et al. (2017), found that visual cues to uncertainty resulted in better performance for both younger and older children, suggesting a bootstrapping effect of facial gesture for the learning of intonation. The same effect was not found for our data. We found that children performed quite similarly for both of the unimodal cues. It was in turn the co-presence of the cues (i.e. having both intonation and facial gesture present) that was most helpful to the participants in our study. That is to say, the more cues to disbelief that were available, the better our participants performed. Our adult controls did not show this effect. Even though both studies investigated epistemic meanings of intonation, the specific meanings of interest in Hübscher et al's study versus our own (uncertainty vs. disbelief) differed, as did the specific tunes (L*H% vs. L*LH%) that encode these meanings. The differences in our findings could be due to these dissimilarities. Cross-linguistic work looking at different types of epistemic meanings and varied tune types could be helpful in understanding to what extent facial gesture plays a bootstrapping role in intonational development. For instance, Crespo-Sendra et al. (2013) found crosslinguistic differences for Dutch- vs. Catalan-speaking adults in terms of the relative weights assigned to facial gesture vs. intonation in comprehension, and therefore children may learn to rely on facial gesture vs. intonation to differing degrees based on the specific form-meaning pairing in the language they are acquiring. Additionally, in our stimuli, since the actors were instructed to produce the stimuli in the most natural way possible, phonation type cues may have also been available in our stimuli for the disbelief condition, providing an extra prosodic cue to the difference in echo question types. Children's use of other prosodic cues like phonation type in comprehending intonational meaning is also an area that merits exploration. On the other hand, both studies show that children develop the ability to use both prosodic and visual cues in tandem, as Hübscher (2018) has also found in production. Hübscher argues that both gestural and prosodic cues lead the way to pragmatic development.

Our study was also novel with respect to studies of L1 prosodic acquisition in that it included a cognitive measure (a false belief task) to assess children's belief reasoning skills. Our hypothesis that children with more sophisticated belief reasoning skills would be better at perceiving disbelief was confirmed: results showed that children who passed our Theory of Mind task were also more successful on our disbelief comprehension task, regardless of condition. The ability to attribute a false belief to another person was predictive for all three conditions, indicating that this ability predicts being able to identify disbelief in others notwithstanding how it is conveyed. Resches & Pérez Pereira (2007) point out that 'the capacity to take into account mental states in others seems to be a key factor which regulates communicative interchanges' (p. 22). They showed that children with higher level ToM abilities were more adept at regulating communication (see also Roby & Kidd, 2008; Sidera et al. 2018; Graham, San Juan & Khu, 2017 for the relationship between Theory of Mind and referential communication) since they were able to understand and anticipate the behaviors of others. While their results were based on production, we might assume that information such as intonation and facial gesture would be some of the cues children might exploit in order to assess and ultimately predict the behavior of another individual in conversation. Thus if a child perceives their interlocutor to be in a state of disbelief, they can decide to address this belief state in their following turn. Resches & Pérez Pereira also discuss the idea that in developing communicative efficiency, children must 1) recognize that others have perspectives different from their own and 2) be able to use this perspective taking as a tool for communication, in turn making relevant inferences on which to base their message. It is quite possible that the children who passed our task, in a real world conversation, might take the information about the belief states they inferred through intonation/gesture, and base their following turn on that information, in effect basing their message on relevant inferences, as Reches & Pérez Pereira suggest. Thus an adult-like response to inferences about disbelief could result in a response such as A's in (2):

(2)
A: I saw an armadillo yesterday.
B: An armadillo?! [produced with disbelief facial gesture]
A: **I couldn't believe it either**!

While our task did not require children to produce any sort of response, it offers a closer look at children's ability to use both intonation and facial gesture as cues to the perspectives of others, which is a crucial piece of pragmatic development, and as stated above this ability is predicted by their ability to attribute a false belief to another individual. San Juan, Khu and Graham (2015) note that by five years of age, children are able to 'rapidly form and integrate perspective inferences to constrain their comprehension of spoken language' (p. 248). Both intonation and facial gesture give rise to such inferences. Our results also confirm that by age five, children are truly becoming quite adept at forming and integrating perspective inferences, not only through spoken language as pointed out by San Juan et al., but also through facial gesture. These authors also note that it is unclear specifically *how* children integrate perspective reasoning and language comprehension. While our results do not speak specifically to this integration, they pinpoint the types of cues that children are using for perspective-taking, and show that the cognitive ability to recognize that the beliefs of individuals differ facilitates the access to the perspectives of others, in this case the epistemic state of another individual. Our results therefore help to provide a more robust picture of why children with more sophisticated ToM skills might be better at regulating communication: children with this profile do better at taking advantage of cues like intonation and facial gesture to gain access to the epistemic states of individuals. This type of access to belief states would of course be paramount for pragmatic development, since it is directly related to observing Gricean Maxims, for example the ability to provide a relevant response (Maxim of Relevance), or the amount of information to provide based on an interlocutor's belief state (Maxim of Quantity).

While we only included one cognitive measure, it will be important in future studies to include a battery of cognitive measures for a more robust snapshot of cognitive abliity. However, studies in prosodic acquisition have not traditionally included such measures. Our study is novel in this sense, and also demonstrates the importance of including cognitive measures to better predict children's performance. Astington and Jenkins (1999) discuss three aspects of language that are related to ToM development, with each playing a different role: pragmatics, semantics and syntax. A child's pragmatic ability is related to ToM by definition, according to these authors, since such an ability entails a child's ability to use language in context, and necessarily includes reasoning about the mental states and intentions of conversational participants. Verbs like *think, know* and *remember* refer to physically unobservable states, and relationships between their acquisition and ToM development have been reported (Olson 1988; Moore, Pure & Furrow, 1990). Papafragou, Fairchild, Cohen and Friedburg (2017) found that the tracking of speakers' mental states is used when acquiring a new word from a person, and that this ability is developing between the ages of 3 and 5. For syntax it has been argued by de Villiers (2007) that the syntax of sentence complements under certain verbs is what facilitates reasoning about the knowledge states of others, claiming that language helps the development of ToM reasoning. de Villiers points out, however, that the influence of ToM development on language development and vice versa is not always so clear, or easy to tease out, especially between the ages of two and four. This is because of (1) either the lack of nonverbal indices being used to explore directionality/correlation with language tasks and (2) a lack of focus on this relationship for children with language delays. However, she notes that specific meanings such as epistemicity and evidentiality present exciting opportunities to explore the relationship between language and ToM. Here we explored the epistemic meaning of disbelief as conveyed through intonational cues and cues on the face. To our knowledge, there have been no studies specifically examining the relationship between ToM development and intonational development using traditional false belief tasks, much less intonational development related to belief states. Our results add to this body of research, suggesting that false belief understanding helps children to comprehend disbelief through different modalities.

Taking our findings together, we can make the following broad conclusions: first, as Nelson (1987) pointed out for facial gesture comprehension in humans, there is a long incubation period for the comprehension of both prosody (in this case a specific intonational melody) of a specific belief state (in this case disbelief) and relevant facial gesture. Our study adds to the existing evidence that between ages 3 and 5, important developments are taking place for children's comprehension of belief states through both intonation and facial gesture, and that by the end of this window more adult-like behavior emerges. Unlike other studies, however, we show no significant difference between intonation-only and facial gesture-only for perceiving belief, and no facilitating effect of facial gesture specifically. This highlights the important role of intonation in a child's understanding of disbelief, and suggests that for some meanings, intonation and facial gesture may develop in tandem with each other. On the other hand, results also show that we can expect children with more sophisticated belief reasoning skills to more readily comprehend disbelief meaning as conveyed through intonation and facial gesture. While the effect of false belief reasoning should be tested with other types of epistemic meaning, we would predict that similar results should be found for other types of mental states that are expressed through intonation and/or facial gesture. The relationship between ToM skills and epistemic meaning, as de Villiers (2007) suggested, has proven to be a useful relationship to explore, and should continue to be explored. Our results also add to the literature on how children are able to comprehend the meaning of facial gesture, in a domain different from, though similar to, emotions. We also leave open the possibility that there is a dynamic relationship between information encoded through intonation and information encoded through facial gesture, such that they mutually influence each other's acquisition, a hypothesis that can be explored in future work. Research on the intonation of different types of belief states and their accompanying prosodic and facial gesture patterns will be important in future research as well. Additional measures such as executive function and working memory, as well as measures of both receptive and expressive language should also be included in future work. To our knowledge, though, this is the first study to measure cognitive factors such as belief reasoning and its role in the acquisition of audiovisual prosody, which has allowed us a more nuanced picture of the acquisition process. Our work reveals the dynamic nature of the factors involved as children learn to "read the minds" of others.

Acknowledgments

This research was funded by a research grant awarded by the Spanish Ministry of Science and Innovation (FFI2015-66533-P, "Intonational and gestural meaning in language"), and by a grant awarded by the Generalitat de Catalunya (2017 SGR 971) to the Prosodic Studies Group. We are thankful to the following schools for their collaboration and participation in this study: CE Jacint Verdaguer (Sant Sadurní d'Anoia), Escola Sant Martí (Arenys de Munt), EP Estalella i Graells (Vilafranca del Penedès) and Farigola del Clot (Barcelona).

References

Armstrong, M.E. (2014). Child comprehension of intonationally-encoded disbelief. In W. Orman & M.J. Valleau (Eds.), *BUCLD 38: Proceedings of the 38th Annual Boston University Conference on Language Development* vol. 2, (pp. 25-38). Somerville, MA: Cascadilla Press.

Armstrong, M.E. (2018). Production of mental state intonation in the speech of toddlers and their caretakers. *Language Acquisition, 25*(2), 119-149.

Astington J.W., Jenkins J.M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology*, *35,* 1311–1320.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21,* 37-46.

Barrera, M.E. & Maurer, D. (1981). The perception of facial expressions by the three-month-old. *Child Development*, *52*, 203-206.

Bavelas, J., Gerwing, J., & Healing, S. (2014). Including facial gestures in gesture-speech ensembles. In M. Seyfeddinjpur & M. Gullberg (Eds.), *From gesture to conversation in visible action as utterance: Essays in honor of Adam Kendon* (pp. 15-34). Amsterdam: John Benjamins Publishing.

Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech. When hand and mouth come together. In D. McNeill (Ed.), *Language and gesture* (pp. 235-257). New York: Cambridge University Press.

Cooper, R.P. & R.N. Aslin. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*(5), 1584-1595.

Crespo-Sendra, V., Kaland, C., Swerts, M., & Prieto, P. (2013). Perceiving incredulity:

the role of intonation and facial gestures. *Journal of Pragmatics*, *47,* 1-13.

de Villiers, J. (2007). The interface of language and Theory of Mind. *Lingua, 117*(11), 1858-1878.

Demir, Ö.E., Fisher, J.A., Goldin-Meadow, S., & Levine, S.C. (2014). Narrative processing in typically developing children and children with early unilateral brain injury: Seeing gesture matters. *Developmental Psychology, 50*(3), 815-828.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Devlopment*, *64*, 657-674.

Friend, M. (2001). The transition from affective to linguistic meaning. *First Language, 21,* 219-243.

Goldman, A.I. (2012). Theory of Mind. In E. Margolis, R. Samuels, & S.P. Sitch (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science.* Oxford: Oxford University Press.

Graham, S.A., San Juan, V., & Khu, M. (2017). Words are not enough: how preschoolers' integration of perspective and emotion informs their referential understanding. *Journal of Child Language, 44*(3), 500-526.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, vol. 3, Speech Acts* (pp. 41-58). New York: Academic Press.

Grossman, R., Striano, T. & Friederici, A.D. (2007). Developmental changes in infants' processing of happy and angry facial expressions: a neurobehavioral study. *Brain Cognition*, *64*(1), 30-41.

Hualde, J.I. & Prieto, P. (2015). Intonational variation in Spanish: European and American varieties. In S. Frota & P. Prieto (Eds.), *Intonational variation in*

Romance (pp. 350-391). Oxford: Oxford University Press.

Hübscher, I. (2018). Preschoolers' pragmatic development: How prosody and gesture lend a helping hand. Unpublished Ph.D. dissertation. Universitat Pompeu Fabra.

Hübscher, I., Esteve-Gibert, N., Igualada, A., & Prieto, P. (2017). Prosody and gesture as bootstrapping devices to pragmatic meaning: How children learn to understand uncertainty. *First Language, 37*(1), 24-41.

Kelly, S. D. (2001). Broadening the units of analysis in communication: speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*(2), 325–349.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.

Klinnert, M.D. (1984). The regulation of infant behavior by maternal facial expression. *Infant Behavior and Development, 7*(4), 447-465.

Kreutzer, M.A. & Charlesworth, W.R. (1973). Infants' reactions to different expressions of emotion. Paper presented at the meeting of the Society for Research in Child Development, Philadelphia, PA.

Ladd, D.R. 2008. *Intonational phonology.* Cambridge: Cambridge University Press.

Mampe, B., Friederici, A.D., Cristophe, A., & Wermke, K. (2009). Newborn's cry melody is shaped by their native language. *Current Biology, 19*(23), 1994-1997.

McNeil, N., Alibali, M., & Evans, J. (2000). The role of gesture in children's comprehension of spoken language: Now they need it now they don't. *Journal of Nonverbal Behavior*, *24*, 131–50.

McNeill, D., Cassell, J., & McCullough, K.E. (1994). Communicative effects of speech mismatched gestures. *Research on Language and Social Interaction*, *27*, 223-237.

Meltzoff, A.N. & Moore, M.K. (1983). Newborn infants imitate adult facial gestures. *Child Development*, *54*, 702-709.

Moore, C., Pure, K. & Furrow, D. (1990). Children's understanding of the modal expressions of speaker certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61,* 722-730.

Moore, C., Harris, L., & Patriquin, M. (2003). Lexical and prosodic cues in the comprehension of relative certainty. *Journal of Child Language, 20*(1), 153–167.

Mumme, D.L., Fernald, A., & Herrera, C. (1996). Infants' responses to facial and vocal emotional signs in social referencing paradigm. *Child Development*, *67*, 3219-3237.

Mundy, P., & Newell, L. (2007). Attention, Joint Attention, and Social Cognition. *Current Directions in Psychological Science*, *16*(5), 269–274.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception & Performance, 24*(3), 756-766.

Nelson, C.A. (1987). The recognition of facial expressions in the first two years of life: mechanisms of development. *Child Development, 58*(4), 889-909.

Nelson, C.A., & Horowitz, F.D. (1983). The perception of facial expressions and stimulus motion by 2- and 5-month-old infants using holographic stimulu. *Child Development, 54,* 868-877.

Nelson, N. & Russell, J.A. (2011). Preschoolers' use of dynamic facial, bodily and vocal cues to emotion. *Journal of Experimental Child Psychology, 110,* 52-61.

Nelson, N. & Russell, J.A. (2016). A facial expression of pax: Assessing children's "recognition" of emotion from faces. *Journal of Experimental Child Psychology, 141,* 49-64.

Nelson, N., Widen, S.C., & Russell, J.A. (2007). *The development of preschooler's Theory of Mind and emotion understanding*. Poster presented at the Bi-Annual Meeting of the Cognitive Development Society, Sante Fe, NM.

Olson, D.R. (1988). On the origins of beliefs and other intentional states in children. In J.W. Astington, P.L. Harris, & D.R. Olson (Eds.), *Developing theories of mind* (pp. 414-426). New York: Cambridge University Press.

Papafragou, A., Fairchild, K., Cohen, M.L., & Friedburg, C. (2017). Learning words from speakers with false beliefs. *Journal of Child Language, 44*(4), 905-923.

Prieto, P. & Rigau, G. (2007). The syntax-prosody interface: Catalan interrogative sentences headed by *que. Journal of Portuguese Linguistics, 6*(2), 29-59.

Prieto, P. (2014). The intonational phonology of Catalan. In S-A. Jun (Ed.), *Prosodic Typology II* (pp. 43-80). Oxford: Oxford University Press.

Resches, M., & Pérez Pereira, M. (2007). Referential communication abilities and Theory of Mind development in preschool children. *Journal of Child Language, 34,* 21-52.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org.

Redcay, E. (2008). The superior temporal sulcus performs a common function for social and speech perception: Implications for the emergence of autism. *Neuroscience and Behavioral Reviews, 32,* 123-142.

Roby, A., & Kidd, E. (2008). The referential communication skills of children with imaginary companions. *Developmental Science, 11*(4), 531-540.

San Juan, V., Khu, M., & Graham, S.A. (2015). A new perspective on children's communicative perspective taking: When and how do children use perspective inferences to inform their comprehension of spoken language? *Child Development Perspectives, 9*(4), 245-249.

Scorce, J.F., Emde, R.N., Campos, J.J., & Klinnert, M.D. (1985). Maternal emotional signaling: its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology, 21*(1), 195-200.

Sidera, F., Perpiñà, G., Serrano, J., & Rostan, C. (2018). Why is Theory of Mind important for referential communication? *Current Psychology, 37*(1), 82-97.

Trevarthen, C., & Hubley, P. (1978). Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. In A. Lock (Eds.), *Action, gesture and symbol: The emergence of language*, (pp. 183-229). New York: Academic Press.

Wermke, K., Teiser, J., Yovsi, E., Kohleberg, P.J, Wermke, P., Robb, M., Keller, H., & Lamm, B. (2016). Fundamental frequency variation within neonatal crying: Does ambient language matter? *Speech, Language and Hearing, 19*(4), 2050-5728.

Widen, S.C., & Russell, J.A. (2008). Children acquire emotion categories gradually. *Cognitive Development, 23,* 291-312.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128.

Young-Brown, G., Rosenfeld, H.M. & Horowitz, F.D. (1977). Infant discrimination of facial expressions. *Child Development, 49,* 555-562.

Appendix A – Stills from adapted false belief task

1a. The princess puts a ball in the right-hand container, and covers it.

1b. The lion moves the ball from the righthand container to the lefthand container.

1c. The princess comes back from school

Table 1. Mixed model results for best fit model – effect of Age in months, Condition and Theory of Mind on Task accuracy. Condition baselines were changed in A, B and C

A.

	Estimate	SE	z value	p value
(Intercept) Baseline=AO, FAIL	-1.35	0.57	-2.37	<0.05
Age in months	0.04	0.01	3.77	<0.001
ToM (PASS)	0.79	0.25	3.18	<0.01
Condition (AV)	1.09	0.28	3.94	<0.001
Condition (VO)	0.17	0.25	0.67	0.50

B.

	Estimate	SE	z value	p value
(Intercept) Baseline=VO, FAIL	-1.18	0.58	-2.07	<0.05
Age in months	0.04	0.01	3.77	<0.001
ToM (PASS)	0.79	0.25	3.18	<0.01
Condition (AO)	-0.17	0.25	-0.67	0.50
Condition (AV)	0.92	0.27	3.37	<0.001

C.

	Estimate	SE	z value	p value
(Intercept) Baseline=AV, FAIL	-0.26	0.58	-0.46	0.65
Age in months	0.04	0.01	3.77	<0.001
ToM (PASS)	0.79	0.25	3.18	<0.01
Condition (AO)	-1.09	0.28	3.94	<0.001
Condition (VO)	-0.92	0.27	-3.37	<0.001

Figure 1: Pitch track, spectrogram and waveform for the disbelief echo question *Una balena?!* ‘A whale? produced with a L*+H prenuclear pitch accent and a L* LH% nuclear configuration in the Cat_ToBI system.

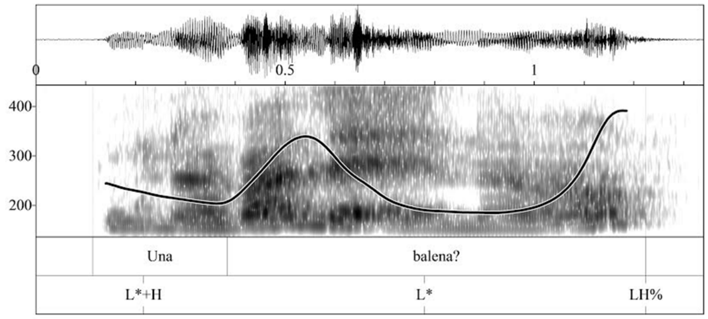

Figure 2: Pitch track, spectrogram and waveform for the neutral echo question *Una balena?* ‘A whale? produced with a L+¡H* L% nuclear configuration.

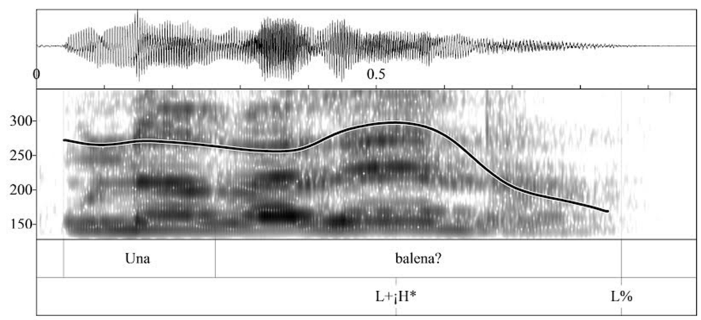

Figure 3. Left panels indicate typical facial gestures for echo questions, produced with brow raising. Right panels indicate typical facial gestures for disbelieving questions, typically produced with backwards movement of the head as well as brow furrowing.

Figure 4. Example of test slide presented to children for the AO condition.

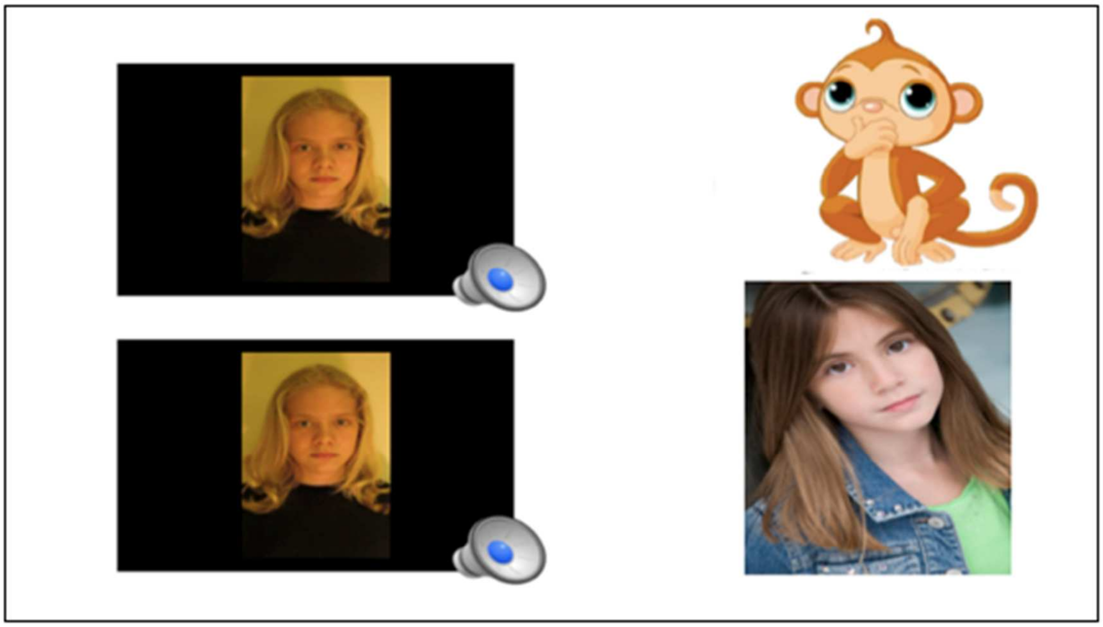

Figure 5. Regression lines for % correct (y axis) by Condition and Age in Months (x axis)