frontiers in Communication

ORIGINAL RESEARCH
published: 18 February 2019
doi: 10.3389/fcomm.2018.00060

The Multimodal Perception of Contrastive Focus in French: A Developmental Study

Lucile Rapin and Lucie Ménard*

Laboratoire de phonétique, Center for Research on Brain, Language, and Music (CRBLM), Université du Québec à Montréal, Montreal, QC, Canada

The current study aimed to better understand the development of prosody perception, by investigating the audiovisual, audio, and visual perception of contrastive focus in French-speaking adults and children. Specifically, 20 adults and 20 school-aged children were presented with short sentences in audiovisual, audio, and visual modalities and were asked to determine if the sentences were produced under neutral or contrastive focused speech. Target words incorporated into the sentences varied across four vowels: /i y u a/. Overall, the adults performed significantly better than the children. Moreover, the children relied more on duration cues to identify contrastive focus, while the adults relied more on formant and lip height values. These findings suggest that children acquire visual cues of speech perception as they mature.

Keywords: contrastive focus, perception, multimodality, development, signal detection theory

OPEN ACCESS

Edited by:
Pia Knoeferle, Humboldt-Universität zu Berlin, Germany

Reviewed by:
Marilyn Vihman, University of York, United Kingdom
Fangfang Li, University of Lethbridge, Canada
Antje Sauermann, Zentrum für Allgemeine Sprachwissenschaft, Canada

****Correspondence:***
Lucile Rapin
lucilerapin@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *23 March 2018*
Accepted: *12 December 2018*
Published: *18 February 2019*

Citation:
Rapin L and Ménard L (2019) The Multimodal Perception of Contrastive Focus in French: A Developmental Study. Front. Commun. 3:60. doi: 10.3389/fcomm.2018.00060

INTRODUCTION

Speech development entails the gradual mastery of orofacial articulators and the refinement of sensory processing. As the child matures, specific movements of the jaw, lips, and tongue are associated with their sensory consequences. It is well-known that those consequences are multimodal. As demonstrated by many studies conducted with adults, auditory, and visual cues are involved in the perception of phonemic units (Robert-Ribes et al., 1998) as well as prosodic prominence (Dohen and Lœvenbruck, 2009). Although audiovisual interaction in early speech perception has been studied in infants, little is known about changes in school-aged children. In this study, we investigated the audiovisual perception of a specific prosodic form, namely contrastive focus, in school-aged French-speaking children and adults.

Prosodic Prominence: The Case of Contrastive Focus

In day-to-day conversation, speech sounds are highly variable. Part of this variability comes from the prosodic structure within which sentences are embedded. Indeed, some sounds are made stronger (more prominent) through various strategies. Prosodic prominence or "narrow focus" refers to emphasis on the unit (word or phrase, for instance) put forward by the speaker in contrast to other units. One function of prosodic prominence is to signal important information in a sentence (e.g., the word *apple* in the sentence "No, I ate the *apple*" in reply to the question "Did you eat the orange?"). Contrastive focus, which is sometimes referred to as "focal accent," "contrastive emphasis," or "contrastive stress," is a type of narrow focus that can be defined as "emphasis on a given constituent of a message that is selected by the speaker, as opposed to emphasizing

another constituent in a paradigmatic comparison" (Selkirk, 1984; Touati, 1987; Pierrehumbert and Hirshberg, 1990; Bartels and Kingston, 1994; Dahan and Bernard, 1996; Ladd, 1996; Di Cristo, 2000). This prosodic form is particularly important in speech development. The present study examines how children learn to perceive contrastive focus in French conveyed by prosody.

Contrastive Focus in Adults

The production of French contrastive focus is well-described in adults. Jun and Fougeron (2000) noted that this prosodic prominence is marked by a considerable rising of pitch contour, which can be assigned to either the first or final syllable of an emphasized (accented) word depending on the length of the word. Acoustically, contrastive focus is produced with increased pitch, intensity, and duration of the accented constituent relative to the others (Dahan and Bernard, 1996; Di Cristo, 1998; Jun and Fougeron, 2000). It is usually accompanied by a hyperarticulation of the accented syllable followed by a hypoarticulation of the subsequent syllable (Lœvenbruck, 1999). Specifically, more important displacements of the jaw, lips, and tongue are observed in syllables produced under contrastive focus by French speakers (Lœvenbruck, 1999; Ménard et al., 2006, 2014).

This hyperarticulation, combined with pre- and post-focus hypoarticulation, should make the focused constituent more perceptually salient (prominent). Studies have indeed revealed that contrastive focus in French is well-perceived in the auditory modality (Gussenhoven, 1983; Dahan and Bernard, 1996; Dohen and Lœvenbruck, 2009). Furthermore, it has been shown that this type of prosodic prominence can also be identified with a better-than-chance accuracy when only seeing a speaker's face (visual modality)[1] (Dohen et al., 2004; Krahmer and Swerts, 2007). Dohen et al. (2004) investigated the visual perception of contrastive focus in reiterated French speech. Participants correctly identified 86% of the utterances, suggesting that adults are sensitive to visual information related to contrastive focus. In a more recent study, the ability to identify focused constituents when presented with the audio, visual, or audiovisual signals was investigated (Dohen and Lœvenbruck, 2009). When asked to determine which part of a sentence had been misunderstood by a speaker (thus identifying the focused constituent), 31 native French speakers correctly identified the focused constituent at a rate of 97.4% in the audiovisual condition, 95.9% in the auditory condition, and 79% in the visual condition (Dohen and Lœvenbruck, 2009). This work, along with that of others (see Krahmer and Swerts, 2007), suggests that visual cues are important in prosody perception and production in general, and especially in contrastive focus perception.

Contrastive Focus in Children

Prosody plays a crucial role in speech acquisition and development (Mehler et al., 1988; Jusczyk and Krumhansl, 1993; Morgan and Demuth, 1996). According to the prosodic bootstrapping hypothesis, intonational and temporal variations in speech provide infants with important cues for word segmentation and for acquiring lexical and morphosyntactic competence (Gleitman et al., 1988; Fernald and Mazzie, 1991). As soon as children begin combining syllables into words, they have to learn how to alternate between weak and strong syllables, signal stressed syllables, and delimit group boundaries. Since focalization involves selecting an item from other items, it is often considered to be a prosodic deictic[2] (Lœvenbruck, 1999). Focalization can be seen as the vocal equivalent of manually pointing, which co-occurs with first word production in one-year-old infants and which can be correlated with later lexical and morphosyntactic development (Bates and Dick, 2002; Volterra et al., 2005). The ability to point, be it manually or prosodically, is a key component of shared attention involved in adult-child interactions. The production and perception of prosodic focus is a prerequisite for typical language development, as suggested by the fact that this function is often altered in children with language disorders (Connaghan and Patel, 2013).

At the word production level, researchers have suggested that children between 2 and 7 years of age use different cues to produce prosodic prominence compared with adults (Allen and Hawkins, 1980; Pollock et al., 1993; Ballard et al., 2012). Children generally rely more on duration and, to a lesser extent, intensity, to indicate stress (although see Connaghan et al., 2001 and Wells et al., 2004). At the articulatory level, Goffman and Malin (1999) showed that children produced longer lip displacements than adults in the context of stressed words. Regarding prosodic focus, it has long been considered that preschool-aged children correctly produce prosodic focus before they can perceive and process prosodic focus as adults do (Cruttenden, 1985; Szendröi, 2004; Hendriks, 2005; Müller et al., 2006). However, a reinterpretation of previous studies led Chen (2010) to suggest that 4–5 year-old children's production of contrastive focus is similar to their comprehension of such focus. However, they do not display adult-like patterns of focus production and perception. A recent study investigating focus marking in English-, French-, and German-speaking children confirmed this hypothesis. It is important to note that the above-mentioned studies did not include any experiment testing the phonetic implementation of focus (Szendröi et al., 2018). Even though children can produce and perceive focused words, do they manipulate the same cues in the speech signal as adults do? Ménard et al. (2006) studied the production of contrastive focus in 4-year-old, 8-year-old, and adult French Canadians. Participants were asked to produce sentences in two prosodic conditions: neutral or contrastive focus. The 4-year-old children only used acoustic strategies related to variations in intensity and pitch, whereas the older children adopted the same acoustic and articulatory labial strategies as the adult speakers. This suggests that the use of acoustic and articulatory features to produce contrastive focus differed between children and adults in these French-speakers, a finding that was replicated in English speakers by Grigos and Patel (2010).

[1] In the context of this study, auditory cues refer to information that can be retrieved from the sound wave and perceived by the ear, whereas visual cues refer to information related to the lip shape and jaw position that is processed by the eye.

[2] In linguistics, a deictic is a word that can be understood thanks to supplementary contextual information.

These differences at the production level were also found in the few studies that have investigated the perception of contrastive focus in children, all in the auditory modality. Wells et al. (2004) observed developmental improvements between 5-year-olds and 14-year-olds. The findings suggested that the ability to perceive contrastive focus improves as a child matures. In agreement with this maturation hypothesis, Cruttenden (1985) observed that children performed significantly worse than adults in identifying contrastive focus. In Wells et al. (2004), contrastive focus perception differed based on age, with younger children performing less well than older children. The authors suggested that children are able to produce specific prosodic intonations before they can interpret the prosodic information of others, and thus prosodic development continues throughout the school years.

Taken together, these studies suggest that the auditory perception of prosodic contrastive focus is still immature in school-aged children. It is not known, however, if this is also true for visual or audiovisual cues. Since it has been reported that children are less sensitive to visual cues in speech perception and seem to rely more on auditory cues in recognizing speech (Massaro, 1984; Dupont et al., 2005), it might be hypothesized that children become more aware of visual cues in the perception of contrastive focus as they grow older. Thus, the current study aimed to better understand the development of contrastive focus perception, by investigating the audiovisual, audio, and visual perception of contrastive focus in children and adults.

METHODS

Participants

Forty subjects (20 adults and 20 children) participated in the study. The adults were six men and 14 women with a mean age of 31.66 ± 9.02 years (range, 20.3–50.4 years). The children were seven boys and 13 girls who had a mean age of 9.42 ± 0.06 years (range, 8.3–10.4 years). Gender distribution was similar in the two groups ($\chi^2(1) = 0.114$, $p = 0.73$). All subjects were native speakers of Canadian French and had no hearing disorder or visual deficit (that could not be corrected by lenses) or any known neurological condition. Informed consent was obtained from the adult participants and the parents of the child participants. The adults received 20 dollars and the children received gifts worth 20 dollars. The university research ethics committee approved the study.

Corpus

A corpus of eight disyllabic words was created. In our previous study (Ménard et al., 2006), we found that when a disyllabic word is produced in focused condition, the first syllable inherits a focused accent with a high pitch rise, whereas the second syllable may or may not be deaccented. Based on this finding, the target words included, in the first syllable, one of the vowels /i y u a/ and one of the consonants /p t k s/. These vowels were selected because they represent the articulatory and acoustic extreme positions of the vocalic triangle (Vorperian and Kent, 2007). The consonants were chosen based on variability in their articulatory mode and their place of articulation; all of them were voiceless. The target words were selected from the Lexicon 3.80 database (http://www.lexique.org) (New et al., 2001) according to the following criteria: high frequency, disyllabic, and identical syllabic structure (/C_1V_1-$C_2V_2C_3$/). The words also had a neutral emotional content. The selected target words were *canard* (/kanaʁ/; 'duck'), *couronne* (/kuʁɔn/; 'crown'), *culotte* (/kylɔt/; 'underwear'), *pilote* (/pilɔt/; 'pilot'), *poussette* (/pusɛt/; 'stroller'), *salade* (/salad/; 'salad'), *sirène* (/siʁɛn/; 'siren'), and *tunnel* (/tynɛl/; 'tunnel'). The target words were then incorporated into a short sentence: ≪ C'est un/une *target word* ≫ (It is a *target word*) to form the corpus. The target word was placed in the final position of the sentence to facilitate elicitation in children, based on pilot experiments. The sentence was produced first (neutral). The experimenter then asked a question introducing an error on the target word. The speaker was asked to repeat the sentence and correct the experimenter (focus).

Experimental Procedure

The corpus was recorded by a French Canadian male adult speaker in both neutral and contrastive focus conditions. He produced six repetitions of each sentence, which resulted in 96 sentences (4 vowels with 2 words for each vowel, and 6 repetitions of 2 conditions). Blue make up was applied on the speaker's lips, to facilitate data extraction on the image, following a method already used in articulatory phonetics and audio-visual studies (Lallouache, 1990; Robert-Ribes et al., 1998). Acoustic and articulatory values associated with the eight target words were extracted at the vowel midpoint for both vowels (V1 and V2) of the target words [using Praat (Boersma, 2001) and Matlab]. The acoustic parameter of pitch was determined by the autocorrelation algorithm. Duration was measured as the time between vowel onset and vowel offset. Intensity was also measured, in dB. The values of the first three formant frequencies were automatically extracted at vowel midpoint using the Linear Predictive Coding (LPC) algorithm. **Figure 1** shows the values of pitch, duration and intensity for the two prosodic conditions (neutral and focus), averaged across words and repetitions. One-way ANOVAs conducted separately for each of the dependent variables (duration, F0, intensity) revealed that position had a significant effect on vowel duration, with V2 being significantly longer than V1 [$F_{(1,188)} = 203.96$; $p < 0.01$].

As suggested by those results, since the target word was in the final position of the sentence, the effects of final lengthening were noticeable: the second syllable was lengthened, but this effect was similar in both prosodic contexts. Moreover, analyses revealed a significant effect of the interaction of vowel position and prosodic condition: for V1 only, duration was longer when produced under contrastive emphasis compared to the neutral condition [$F_{(1,188)} = 15.28$; $p < 0.05$]. A similar pattern was observed for average F0 values (**Figure 1**, upper right panel). F0 was significantly lower in V2 than in V1 [$F_{(1,188)} = 187.43$; $p < 0.001$], and V1 was produced with a higher F0 in the emphasis condition compared to the neutral condition [$F_{(1,188)} = 29.42$; $p < 0.01$]. Regarding average intensity (**Figure 1**, lower left panel), the ANOVA showed a significant effect of position, with intensity values being significantly larger for V1 than for V2 [$F_{(1,188)} = 312.13$; $p < 0.001$]. Prosodic condition also had

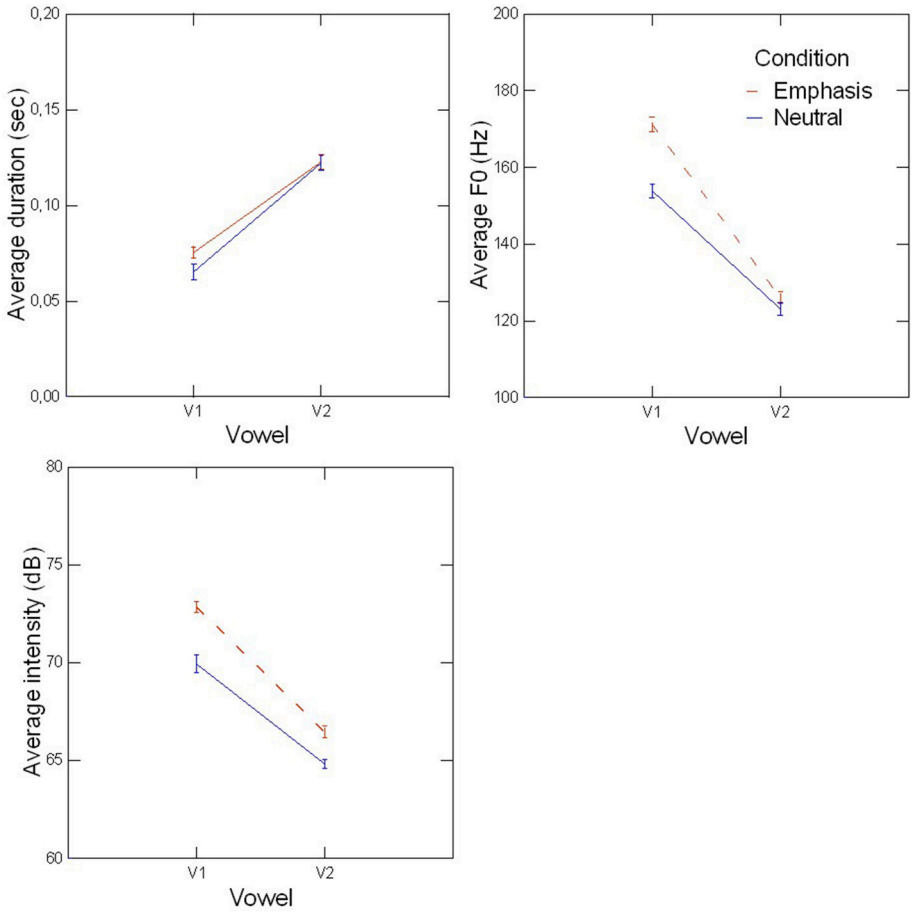

FIGURE 1 | Average duration **(Upper Left)**, F0 **(Upper Right)**, and instensity **(Lower Left)** measured on the first (V1) and second (V2) vowels of the disyllabic target word, for each prosodic condition (neutral and emphasis). Error bars are standard errors.

a significant effect on intensity: in the emphasis condition, values were produced louder than in the neutral condition [$F_{(1,188)} = 48.99$; $p < 0.001$]. No effect of the interaction between position and condition was found for this variable. This agrees with the pattern of results produced by adult participants in our previous study (Ménard et al., 2006).

The articulatory values of lip height and lip width at vowel midpoint were measured manually using Matlab. On each image extracted at vowel midpoint, a trained research assistant selected 4 points on the internal contour of the lips: on the vertical dimension, the maximal and minimal points, and on the horizontal dimension, the leftmost and rightmost points. Values are presented in **Figure 2**, for both vowel positions and prosodic conditions. Lip area was calculated using the following formula: Lip_area= π*Height/2*Width/2. A two-way ANOVA revealed a significant effect of the interaction of vowel and prosodic condition on lip area values [$F_{(3,376)} = 135.64$; $p < 0.05$]: for /a/ and /i/, when produced in emphasis condition, lip area values were larger than in the neutral condition; conversely, for /u/ and /y/, values of lip area were smaller in the emphasis than in the neutral condition. [$F_{(3,376)} = 256.97$; $p < 0.05$]. Lip area values measured on the second vowel V2 (**Figure 2**, right panel) did not differ significantly across prosodic conditions and vowels. Those articulatory correlates of prosodic condition in V1 affected formant frequencies, as shown in **Figure 3**. In agreement with the articulatory results, only V1 was significantly affected by prosodic condition [significant effect of the interaction between condition and position: $F_{(3,376)} = 243.52$; $p < 0.05$]. For the vowels /a/ and /i/, F1 was higher in the emphasis condition than in the neutral condition [$F_{(3,376)} = 185.42$; $p < 0.05$] whereas for /y/ and /u/, F1 was smaller in the emphasis condition than in the neutral condition [$F_{(3,376)} = 204.52$; $p < 0.05$]. In summary, focus had clear acoustic and articulatory correlates, mainly affecting the first syllable of the disyllabic target word.

Images of the lower part of the face (from just under the chin to the middle of the nose) were extracted and then implemented in PsychoPy software (Peirce, 2009). The image of the face took up a third of the height of the computer screen display and was centered on the screen. The sentences were presented in a random order in three modalities: audiovisual, visual, and audio. In the audio modality, the participant could only hear the sound. In the visual modality, the participant could only see the lower

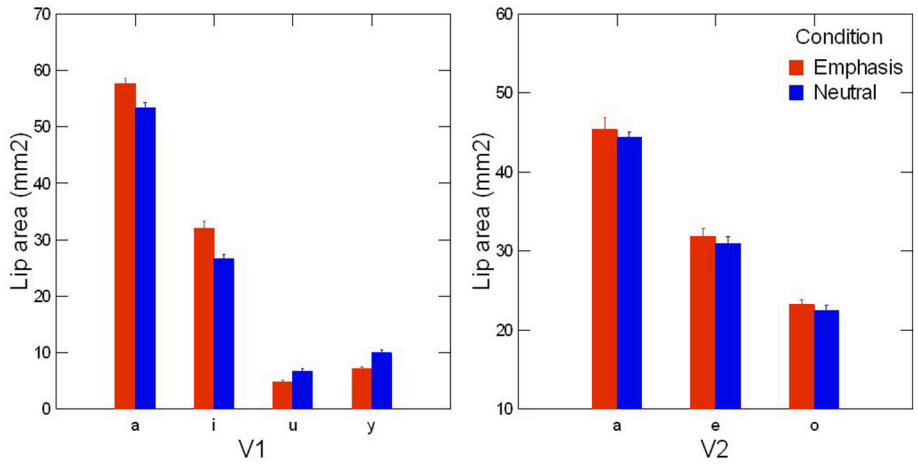

FIGURE 2 | Average lip area values measured on the first vowel V1 **(Left)** and the second vowel V2 **(Right)** of the target word, for each prosodic condition (neutral and contrastive focus). Error bars are standard errors.

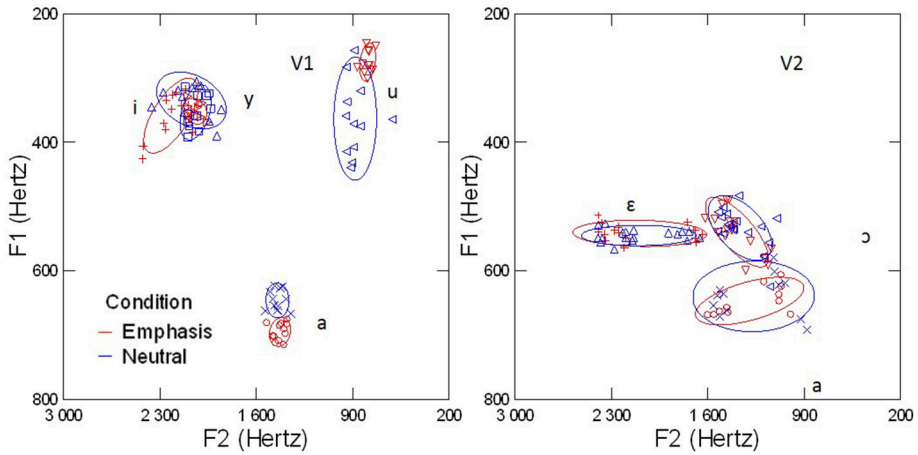

FIGURE 3 | F1 and F2 values for the first vowel of the target word (V1, **Left**) and for the second vowel of the target word (V2, **Right**), for contrastive focus and neutral conditions. Ellipses enclose 95% of the dispersion.

part of the speaker's face during the production of the sentence. An example of a still image that was part of the visual display presented to the participant is provided in **Figure 4**. In the audio-visual condition, both the images and the sound were presented to the participant. Each modality contained the 96 trials.

Participants sat in front of a laptop computer in a quiet room. The subjects had to determine if sentences produced in the audiovisual, visual, and audio modalities were produced in neutral speech or contrastive focus and indicate this by clicking on "N" or "F" keys on the keyboard that corresponded with the "neutral" or "focus" conditions. A familiarization task was included in the test. The experimenter provided children with examples of focused and neutral conditions (items that were not included in the test). The experimenter would start the experiment only when the child acknowledged that she/he understood the task and was ready to perform it. This kind of task was also used with preschool-aged children in Szendröi et al. (2018) study and proved to be reliable. Each child was supervised by an experimenter, who made sure the child remained focused on the task. The perceived condition (neutral or focus) as well as reaction time were collected.

Data Analyses

Statistical tests were conducted using IBM SPSS Statistics, version 22.0 (IBM Corp, 2013). Using signal detection theory methods, the data were used to calculate a sensitivity index d prime: d'= [z(hit) – z(false alarm)] (Macmillan and Creelman, 2004). This index measures the distance between the signal and noise mean distributions and takes into account the tendency to respond "neutral" or "focus" (Stanislaw and Todorov, 1999). In the current study, the signal was the focused speech and the noise was the neutral speech.

FIGURE 4 | Example of a visual stimulus presented to the participant.

Because some adult participants had extreme values of 0 or 1 in their hit rates and false alarm rates (i.e., ceiling effects), we used the log-linear approach proposed by Stanislaw and Todorov (1999) to overcome this. As they explain, this approach " involves adding 0.5 to both the number of hits and the number of false alarms and adding 1 to both the number of signal trials and the number of noise trials, before calculating the hit and false-alarm rates." We computed d' scores for each subject for each modality and vowel combination. A d' value of 0 corresponded to an inability to distinguish between signal (focus) and noise (neutral) conditions, whereas larger values indicated a better capacity to distinguish focus from neutral speech. Reaction times were also extracted.

For each dependent variable (d' and reaction time), a mixed ANOVA was performed with modality (audiovisual [AV] vs. audio [A] vs. visual [V]) and vowel (/a/ vs. /i/ vs. /u/ vs. /y/) as within-subject factors and group (adults vs. children) as the between-subject factor. Interaction effects and pairwise comparisons were then explored using the Bonferroni correction with the alpha level set to 0.05. Effect sizes were calculated as partial eta squared (η_p^2). Only significant main effects and interactions are reported. Finally, in order to investigate the weight of the various acoustic and articulatory cues in the identification of focus, multiple regression analyses were conducted for each age group, with perceptual responses as the dependent variable, and acoustic parameters (duration, pitch, intensity, formants) and articulatory parameters (lip area) as the independent variables.

RESULTS

The perception experiment results were well-above chance in both groups (adults and children), with an overall identification rate of 76.9% for adults and 66.1% for children (**Table 1**). One-sample Student *t*-tests for above-chance identification indicated that the visual perception of contrastive focus significantly exceeded chance for both adults $t_{(19)} = 5.81$, $p < 0.001$ and children $t_{(19)} = 2.4$, $p < 0.027$.

TABLE 1 | Identification rates for the perception of contrastive focus test, for different modalities, in adult, and child participants.

Identification rate	Total (%)	AV (%)	A (%)	V (%)
Adults	76.89	86.04	89.11	55.2
Children	66.08	72.34	73.9	52.0

AV, audiovisual; A, audio; V, visual.

Average D-Prime Values

A Mauchly's sphericity test indicated that the assumption of sphericity was not violated for any of the independent variables. The mixed ANOVA revealed a main effect of modality [$F_{(2,76)} = 246.63$, $p < 0.001$, $\eta_p^2 = 0.87$]. Pairwise comparisons showed that average d' scores in the audiovisual and audio modalities did not differ (1.76 and 1.9, respectively), whereas they both differed significantly from the visual modality (0.22), each with $p < 0.001$. Thus, the ability to distinguish between contrastive focus and neutral speech was more difficult when the stimuli were presented in the visual modality only, for both children and adults. Vowels also had a significant main effect [$F_{(3,114)} = 14.32$, $p < 0.001$, $\eta_p^2 = 0.27$], suggesting that there were different mean d' scores, depending on the vowel. Pairwise comparisons revealed that the vowels /a/ and /y/ led to significantly higher d' scores (1.47 and 1.41, respectively) than the vowels /u/ (1.23) and /i/ (1.07), each with $p < 0.003$. The group main effect was statistically significant [$F_{(1,38)} = 114.45$, $p < 0.001$, $\eta_p^2 = 0.75$]. Adult participants performed considerably better (1.69) than child participants (0.9). That is, adults were better able to distinguish between contrastive focus and neutral speech than children, in all modalities.

There were also interaction effects. Modality and vowels interacted such that score patterns were similar across vowels for the audiovisual and audio modalities but not for the visual modality [$F_{(6,228)} = 10.63$, $p < 0.001$, $\eta_p^2 = 0.22$]. Specifically, for the audiovisual and audio modalities, vowels /a/ (AV = 2, A = 2.1) and /y/ (AV = 1.9, A = 2 .1) elicited the highest d' scores and were significantly different from vowels /u/ (AV = 1.6, A = 1.8) and /i/ (AV = 1.4, A = 1.4) ($p < 0.04$ for all), whereas for the visual modality, the vowel /i/ (0.4) had the highest d' score and was statistically different from the vowels /u/ (0.15) and /y/ (0.1) ($p < 0.04$ for all) (**Figure 5**).

The interaction between the modality and the group was also statistically significant, $F_{(2,76)} = 20.55$, $p < 0.001$, $\eta_p^2 = 0.35$. Specifically, all three modalities led to considerable recognition-ability differences between the children and adults [evidenced by Student t-tests performed on each modality–audiovisual: $t_{(38)} = -7.15$, $p < 0.001$; audio: $t_{(38)} = -8.66$, $p < 0.001$; visual: $t_{(38)} = -2.33$, $p < 0.025$]–but the difference was actually smaller for the visual modality, as illustrated in **Figure 6**.

Reaction Time

For reaction time measures (**Figure 7**), Mauchly's test indicated that the assumption of sphericity had been violated for modality ($\chi^2(2) = 19.94$, $p < 0.001$) and vowel ($\chi^2(5) = 15.16$, $p < 0.01$)

as well as for their interaction, $\chi^2(20) = 32.47$, $p < 0.039$. The Greenhouse-Geisser corrected-tests were consequently reported for the effects linked with these variables and their interactions. The mixed ANOVA revealed that the modality had a significant main effect on reaction time, $F_{(1.41, 53.65)} = 68.19$, $p < 0.001$, $\eta_p^2 = 0.64$. Pairwise comparisons showed that the audiovisual modality resulted in the shortest reaction time (0.96 s), followed by the audio modality (1.32 s), followed by the visual modality (1.51 s), all $p < 0.001$. The group effect was significant, $F_{(1,38)} = 5.28$, $p < 0.027$, $\eta_p^2 = 0.12$. Overall, adults had considerably shorter reaction times (1.12 s) than the children (1.41 s). This is typical of language development, children having overall slower processing mechanisms than adults in perceptual tasks.

Relationships Between Acoustic/Articulatory Parameters and Identification Rates

Overall, adults performed better than children in the perceptual task (as shown by their d' values and reaction times). Both groups performed better in the audiovisual and audio modalities than in the visual modality. To investigate if children used different strategies to identify prosodic focus, we conducted linear mixed effects logistic regressions to determine the role of the different cues related to the perception of focus in children and adults. Perceptual responses were considered the dependent variable, with the perception of the neutral condition being coded as "0," whereas the perception of focus was coded as "1." Apart from speaker group (ref = children), the following dependent variables were considered, based on the significant effect of prosodic condition on their production in the corpus (cf. section Experimental Procedure): pitch, duration, intensity, first formant, and lip area. We performed three independent analyses: (i) in the audio modality, speaker group, pitch, duration, intensity, and first formant were included as fixed effects, and subject-specific intercepts were included as random factors; (ii) in the visual modality, group and lip area values were the fixed effects and subject-specific intercepts were included as random factors; and (iii) in the audiovisual modality, group, pitch, duration,

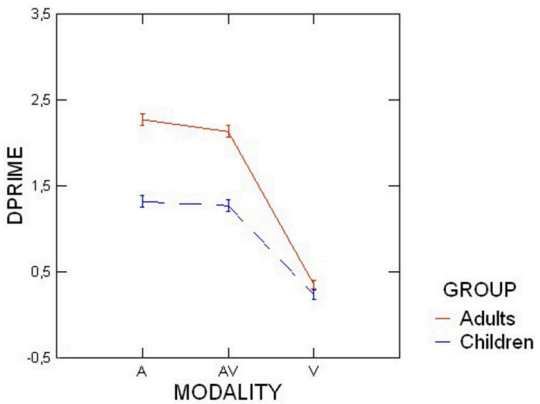

FIGURE 5 | Average d' scores for adult and child participants, for different modalities. AV, audiovisual; A, audio; V, visual. Error bars are standard errors.

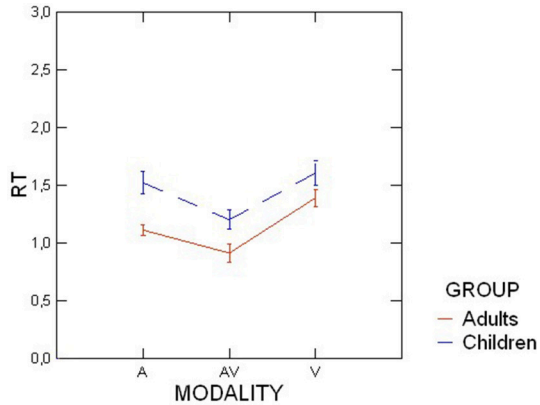

FIGURE 7 | Average values of reaction time for child and adult participants, for each modality and vowel. AV, audiovisual; A, audio; V, visual. Error bars are standard errors.

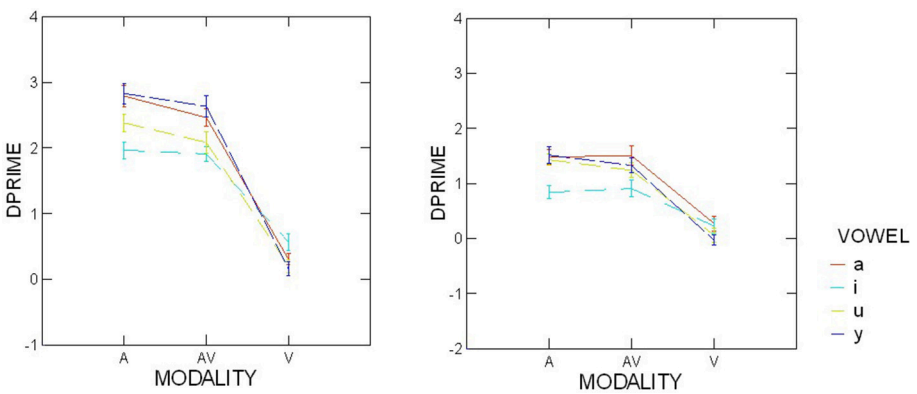

FIGURE 6 | Average d' scores for adults **(Left)** and children **(Right)**, for each modality and vowel. AV, audiovisual; A, audio; V, visual. Error bars are standard errors.

intensity, and lip area were the fixed effects and subject-specific intercepts were included as random factors. Odds ratios are shown in **Table 2**[3].

As can be seen in **Table 2A**, in the audio modality, odds ratios associated with main effects suggest that adults are more likely to perceive focused words than children. Furthermore, results presented in **Table 2A** reveal that vowels with increased values of pitch, duration, and formant values have significantly more chances of being perceived focused. The interaction of speaker group with F1 is significant: the perceived focused words are more often associated with increased F1 and perceived by adult participants than by children. Turning now to the visual modality (**Table 2B**), odds ratios suggest that lip area is a significant predictor of perceived focus: increased lip area is more likely identified as focused. Again, the interaction between lip area and speaker group is significant in the model: increased lip area in the adult group has significantly more chances than in the child group of being perceived as indicating focus. Regarding the audiovisual modality (**Table 2C**), this time, parameters of pitch and duration are not equivalent: words with increased pitch values have less chance of being perceived as focused by adults than by children. However, the odds of perceiving a focused constituent when lip area is enlarged are greater in adults than in children.

TABLE 2 | Results of linear mixed effects logistic regressions performed on perceptual responses (for the group variable, ref = children).

Fixed effects	Odd ratios	Estimate
(A) AUDIO MODALITY		
(Intercept)	5.77	−25.88*
group	4.00	22.11*
Pitch	7.19	20.39*
Duration	5.82	1.76*
F1	5.40	0.62*
group*pitch	0.33	17.42
group*duration	0.41	14.38
group*F1	1.48	0.16*
(B) VISUAL MODALITY		
(Intercept)	1.85	9.73*
group	3.02	2.05*
Lip area	9.88	4.05*
group*lip area	1.02	1.98*
(C) AUDIO-VISUAL MODALITY		
(Intercept)	1.18	1.72*
group	7.26	2.26
Pitch	7.26	0.89*
Duration	9.10	0.22*
Lip area	9.92	3**
group*Pitch	0.69	1.02**
group*duration	1.72	0.30
group*lip area	1.30	4**

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

DISCUSSION

In this investigation of the perception of contrastive focus in 20 adults and 20 school-aged children, contrastive focus was implemented in words combining four vowels (/i y u a/) across three perceptual modalities (audiovisual, audio, and visual). We used an index of sensitivity (d') derived from signal-detection theory to investigate the ability to discriminate between speech with contrastive focus and neutral speech, and we analyzed reaction times.

The results suggest that the modality of stimulus presentation affects the ability to identify contrastive focus. Audiovisual and audio modalities of stimulus presentation were associated with the highest contrastive focus sensitivity scores and shorter reaction times. In the visual modality, participants had noticeable difficulty distinguishing between the two types of speech.

Taken together, these findings highlight the primary role of acoustic cues in the perception of contrastive focus. The results are consistent with previous reports of an efficient auditory-only identification of contrastive focus (Dohen et al., 2004). However, the audiovisual integration was superior, since it resulted in the highest contrastive focus identification scores with the shortest reaction times, which provides evidence that adding visual cues to audio cues improves contrastive focus perception. The integration of vision with hearing may reduce the duration of cognitive processing of the prosodic information and therefore lead to a more accurate and rapid designation of focus (Dohen et al., 2004).

The choice of vowels also had an impact on the correct identification of contrastive focus. The vowels /a/ and /y/ elicited considerably higher recognition scores compared with the vowels /i/ and /u/, for both children and adults, in the audiovisual and audio modalities. Moreover, in the visual modality, the vowel /i/ elicited a significantly higher recognition score relative to /u/ and /y/, which actually was associated with the lowest d' score. The vowels /u/ and /y/ are associated with lip-rounding gestures, which are more visible when a profile is viewed. The fact that the face was displayed from the front and not as a profile in this study may explain why those vowels were not as clearly detected in the visual modality as the vowels /i/ and /a/, for which articulatory gestures are strongly apparent when viewed from the front.

The findings related to the development of contrastive focus identification are of special interest. Overall, adults performed better than children in this experiment, both in terms of perceptual scores than reaction times[4]. Specifically, adults had significantly higher contrastive focus recognition scores than children in the audiovisual and audio modalities of presentation, and to a lesser extent, in the visual modality. These findings add to previous findings that visual cues play a smaller role than acoustic markers in audiovisual speech perception (Massaro, 1984), by adding a prosodic component. Regardless of their age, participants had a harder time understanding prosody in the

[3]The syntax used in R was the following: m <- glmer(rep ~ group + $Ratio_{F0}$ + $Ratio_{duree}$ + $Ratio_{F1}$ + (1 | sujet), data = data, family = binomial, control = glmerControl (optimizer = "bobyqa"), nAGQ = 10).

[4]One might suggest that longer reaction times in children reflect the difficulty of the task. However, a similar task was used by Szendröi et al. (2018) with 3- to 6-year-old children (younger than the child participants in this study), and proved to be appropriate for this population.

visual modality than in the audio modality. It has to be noted that, apart from maturation, vocabulary size might also play a role in the children's poorer performance than adults. Indeed, it has been reported that for very young children or infants, age is less important than vocabulary size as a predictor of speed of response in word recognition (e.g., Fernald et al., 2001) and in fast mapping (Torkildsen et al., 2008, 2009), for example. In a further study, vocabulary size could also be measured and included in the statistical models.

The current investigation of cues that speakers rely on to identify contrastive emphasis showed that children do not rely on segmental information, be it in the auditory (F1) or visual (lip area) modality, to the same extent as adults. The findings can be interpreted in light of the hypothesis that, until 8-9 years of age, children have not learned to process and interpret visual sources of speech information, in either the production or the perception domain, as suggested by Ménard et al. (2006). That study examined 8-year-old children (similar to the current cohort of 9-year-old children) and reported that they did not use lip-opening gestures to mark contrastive focus as much as adults do and instead relied on acoustic features. Repeating the current study in adolescents would likely provide more evidence of the link between the maturation of articulatory gestures and visual correlates of perception.

This developmental profile in speech perception parallels that proposal for speech production in our previous work (Ménard et al., 2006, 2014). Indeed, we found that children first learn the hyperarticulated form of a phoneme, and then have to learn to hypoarticulate, for example, through reduced magnitudes of articulatory displacements. Those results are consistent with those presented in Allen and Hawkins (1980) and Payne et al. (2012). Thus, when we investigated the produced correlates of prosodic focus, we found that 4-year-olds did not differentiate their lip configurations according to the prosodic condition, but instead used pitch, intensity, and duration to signal emphasis. Eight-year-olds showed some differentiation of lip movements according to prosodic context, but to a lesser extent than adults. Taken together, these earlier results of production studies and findings from the present study suggest that there is a close relationship between production and perception in speech development, and children are perceptually attuned to the parameters they use to produce a given form. This is consistent with the articulatory filter hypothesis proposed by Vihman (1993, 2014), in which infants' vocal repertoires influence their ability to retain word forms. A similar phenomenon could be at stake at a later stage of speech development, when children are fine-tuning their speech production and perception abilities.

In conclusion, the investigation of contrastive focus perception in adults and school-aged children suggests that children tend to acquire this ability as they mature. Overall, adults performed better in discriminating contrastive focus. These findings build on previous research of the influence of visual information in speech perception. Although perceiving contrastive focus appeared to be more difficult in the visual modality than in the audio or audiovisual modalities, all participants were still able to do this; however, adults performed better than children, suggesting that visual cues in speech perception are acquired with age, as children learn to use their speech articulators appropriately. This hypothesis could seem to contradict the fact that phonemes involving the lips (such as /p/ and /m/) are acquired early in life. We believe that the weight given to the auditory and visual modality vary throughout development. Early in life, children use visual cues to acquire articulatory control of phonemic goals but as they improve their language and sensory processing ability, they start using the auditory modality more importantly to build feed-forward models of the articulatory-to-acoustic relationships. Thus, when processing complex speech tasks such as the perception of prosodic focus, they rely more on auditory feedback than visual feedback. This interpretation is in line with the results of McGurk tasks in children and adults, showing that children put more weight on visual cues than adults do (Massaro, 1984; Dupont et al., 2005, for instance).

AUTHOR CONTRIBUTIONS

LR and LM designed the study. LR performed the study and analyzed the data. LR and LM wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by an Insight Grant from the Natural Sciences and Engineering Research Council of Canada (grant number 312395-2010) and the Social Sciences and Humanities Research Council of Canada (grant number 430-2012-0659). We thank Marlene Busko for copyediting the paper.

REFERENCES

Allen, G. D., and Hawkins, S. (1980). "Phonological rhythm: definition and development," in *Child Phonology, Vol. 1: Production*, G. Yeni-Komshian, J. F. Kavanagh, and C. A. Ferguson (New York, NY: Academic Press), 227–256.

Ballard, K. J., Djaja, D., Arciuli, J., James, D. G. H., and van Doorn, J. (2012). Developmental trajectory for production of prosody: lexical stress contrastivity in children ages 3 to 7 years and in adults. *J. Speech Lang. Hear. Res.* 55, 1822–1835. doi: 10.1044/1092-4388(2012/11-0257)

Bartels, C., and Kingston, J. (1994). "Salient pitch cues in the perception of contrastive focus," in *Proceedings of the Conference on Focus and Natural Language Processing,* eds P. Bosch and R. Van Der Sandt (Heidelberg: IBM).

Bates, E., and Dick, F. (2002). Language, gesture, and the developing brain. *Dev. Psychobiol.* 40, 293–310. doi: 10.1002/dev.10034

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot Int.* 5, 341–345.

Chen, A. (2010). Is there really an asymmetry in the acquisition of the focus-to-accentuation mapping? *Lingua* 120, 1926–1939. doi: 10.1016/j.lingua.2010.02.012

Connaghan, K. P., Moore, C. A., Reilly, K. J., Almand, K. B., and Steeve, R. W. (2001). *Acoustic and Physiologic Correlates of Stress Production Across Systems,*

Poster Presented at the American Speech-Language-Hearing Association (New Orleans, LA).

Connaghan, K. P., and Patel, R. (2013). Impact of prosodic strategies on vowel intelligiblity in childhood motor speech impairment. *J. Med. Speech Lang. Pathol.* 20, 133–139.

Cruttenden, A. (1985). Intonation comprehension in ten-year-olds. *J. Child Lang.* 12, 643–661. doi: 10.1017/S030500090000670X

Dahan, D., and Bernard, J.-M. (1996). Interspeaker variability in emphatic accent production in French. *Lang. Speech* 39, 341–374. doi: 10.1177/002383099603900402

Di Cristo, A. (1998). "Intonation in French," in *Intonation Systems: A Survey of Twenty Languages,* eds D. J. Hirst and A. Di Cristo (Cambridge, UK: Cambridge University Press).

Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français (deuxième partie). *J. French Lang. Stud.* 10, 27–44. doi: 10.1017/S0959269500004671

Dohen, M., and Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Lang. Speech* 52, 177–206. doi: 10.1177/0023830909103166

Dohen, M., Lœvenbruck, H., Cathiard, M.-A., and Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Commun.* 44, 155–172. doi: 10.1016/j.specom.2004.10.009

Dupont, S., Aubin, J., and Ménard, L. (2005). A study of the mcgurk effect in 4- and 5-year-old french canadian children. *ZAS Papers Linguistics.* 40, 1–17. doi: 10.1.1.603.9384

Fernald, A., and Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Dev. Psychol.* 27, 209–221. doi: 10.1037/0012-1649.27.2.209

Fernald, A., McRoberts, G. W., and Swingley, D. (2001). "Infants' developing competence in recognizing and understanding words in fluent speech," in *Approaches to Bootstrapping. Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition,* eds J. Weissenborn and B. Höhle (Amsterdam; Philadepphia, PA: John Benjamins Publishing Company), 97–123. doi: 10.1075/lald.23.08fer

Gleitman, L., Gleitman, H., Landau, B., and Wanner, E. (1988). "Where learning begins: initial representations for language learning," in *Linguistics: The Cambridge Survey*, Vol. 3, *Language: Psychological and Biological Aspects*, ed F. J. Newmeyer (New York, NY: Cambridge University Press), 150–193.

Goffman, L., and Malin, C. (1999). Metrical effects on speech movements in children and adults. *J. Speech Lang. Hear. Res.* 42, 1003–1115. doi: 10.1044/jslhr.4204.1003

Grigos, M. I., and Patel, R. (2010). Acquisition of articulatory control for sentential focus in children. *J. Phonet.* 38, 706–715. doi: 10.1016/j.wocn.2010.10.005

Gussenhoven, C. (1983). *A Semantic Analysis of the Nuclear Tones of English.* Bloomington, IN: Indiana University Linguistics Club.

Hendriks, P. (2005). "Asymmetries in the acquisition of contrastive stress," in *Paper Presented at the Workshop on Contrast, Information Structure and Intonation*. Stockholm: Stockholm University, (Accessed October 28–29, 2005).

IBM Corp (2013). *IBM SPSS Statistics for Windows*, Version 22.0. Armonk, NY: IBM Corp.

Jun, S.-A., and Fougeron, C. (2000). "A phonological model of French intonation," in *Intonation: Analysis, Modelling and Technology*, ed A. Botinis (Boston, MA: Kluwar Academic Publishers), 209–242.

Jusczyk, P. W., and Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *J. Exp. Psychol. Hum. Percept. Perform*. 19, 627–640.

Krahmer, E., and Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *J. Mem. Lang.* 57, 396–414. doi: 10.1016/j.jml.2007.06.005

Ladd, R. D. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lallouache, T. (1990). "Un poste "visage-parole": acquisition e traitement de contours labiaux," in *Proceedings of the XVIII Journées d'Études sur la Parole, Montréal*, 282–286.

Lœvenbruck, H. (1999). "An investigation of articulatory correlates of the accentual phrase in French," in *Proceedings of the 14th ICPhS*, Vol 1, 667–670. San Francisco, CA.

Macmillan, N. A., and Creelman, C. D. (2004). *Detection theory: A User's Guide.* New York, NY: Psychology Press

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Dev.* 55, 1777–1788. doi: 10.2307/1129925

Mehler, J., Jusczyk, P. W., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition* 29, 143–178. doi: 10.1016/0010-0277(88)90035-2

Ménard, L., Leclerc, A., and Tiede, M. (2014). Articulatory and acoustic correlates of contrastive focus in congenitally blind adults and sighted adults. *J. Speech Lang. Hear. Res.* 57, 793–804. doi: 10.1044/2014_JSLHR-S-12-0395

Ménard, L., Loevenbruck, H., and Savariaux, C. (2006). "Articulatory and acoustic correlates of contrastive focus in French children and adults," in *Speech Production: Models, Phonetic Processes and Techniques*, eds J. Harrington and M. Tabain (New York, NY: Psychology Press), 227–251.

Morgan, J., and Demuth, K. (1996). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Müller, A., Höhle, B., Schmitz, M., and Weissenborn, J. (2006). "Focus-to-stress Alignment in 4- to 5-year-old German-learning Children," in *Proceedings of GALA 2005*, eds A. Belletti, E. Bennati, C. Chesi, E. Di Domenico, I. Ferrari (Cambridge, Cambridge Scholars Publishing), 393–407.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE^{TM}/A lexical database for contemporary French: LEXIQUE^{TM}. *L'année Psychol.* 101, 447–462. doi: 10.3406/psy.2001.1341

Payne, E., Post, B., Astruc, L., Prieto, P., and del Mar Vanrell, M. (2012). Measuring child rhythm. *Lang. Speech* 55, 203–229. doi: 10.1177/0023830911417687

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Front. Neuroinform.* 2:10. doi: 10.3389/neuro.11.010.2008

Pierrehumbert, J., and Hirshberg, J. (1990). "The meaning of intonational contours in discourse," in *Intentions in Communication*, eds P. Cohen, J. Morgan, and M. Pollack (Cambridge: The MIT Press), 271–311.

Pollock, K. E., Brammer, D. M., and Hageman, C. F. (1993). An acoustic analysis of young children's productions of word stress. *J. Phon.* 21, 183–203.

Robert-Ribes, J., Schwartz, J.-L., Lallouache, T., and Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *J. Acoust. Soc. Am.* 103, 3677–3689. doi: 10.1121/1.423069

Selkirk, E. O. (1984). "The grammar of intonation," in *Phonology and Syntax: The Relation Between Sound and Structure*, ed E. O. Selkirk (Cambridge: The MIT Press), 197–296.

Stanislaw, H., and Todorov, N. (1999). Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput.* 31, 137–149. doi: 10.3758/BF03207704

Szendröi, K. (2004). Acquisition evidence for an interface theory of focus. *LOT Occasion. Series* 3, 457–468.

Szendröi, K., Bernard, C., Berger, F., Gervain, J., and Höhle, B. (2018). Acquisition of prosodic focus marking by English, French, and German three-, four-, five- and six-year-olds. *J. Child Lang.* 45, 219–241. doi: 10.1017/S0305000917000071

Torkildsen, J. V. K., Hansen, H. F., Svangstu, J. M., Smith, L., Simonsen, H. G., Moen, I., et al. (2009). Brain dynamics of word familiarization in 20-month-olds: effects of productive vocabulary size. *Brain Langu.* 108, 73–88. doi: 10.1016/j.bandl.2008.09.005

Torkildsen, J. V. K., Svangstu, J. M., Hansen, H. F., Smith, L., and Simonsen, H. G. (2008). Productive vocabulary size predicts event-related potential correlates of fast mapping in 20-month-olds. *J. Cogn. Neurosci.* 20, 1266–1282. doi: 10.1162/jocn.2008.20087

Touati, P. (1987). *Structures Prosodiques du Suédois et du Français, Working Paper 21*. Lund University Press.

Vihman, M. M. (1993). Variable paths to early word production. *J. Phon.* 21, 61–82.

Vihman, M. M. (2014). *Phonological Development: The First Two Years, 2nd Edn*. Malden, MA: Wiley-Blackwell.

Volterra, V., Caselli, M. C., Capirci, O., and Pizzuto, E. (2005). "Gesture and the emergence and development of language," in *Beyond Nature-Nurture: Essays in Honor of Elizabeth Bates*, eds M. Tomasello and D. Slobin (Mahwah, NJ: Lawrence Erlbaum Associates), 3–40.

Vorperian, H. K., and Kent, R. D. (2007). Vowel acoustic space development in children: a synthesis of acoustic and anatomic data. *J. Speech Lang. Hear. Res.* 50, 1510–1545. doi: 10.1044/1092-4388 (2007/104)

Wells, B., Peppé, S., and Goulandris, N. (2004). Intonation development from five to thirteen. *J. Child Lang.* 31, 749–778. doi: 10.1017/S030500090400652X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.