The role of production abilities in the perception of consonant category in infants

Anne Vilain, Marjorie M. Dole, Hélène Loevenbruck, Olivier Pascalis, Jean-Luc Schwartz

(i) Title

The role of production abilities in the perception of consonant category in infants

(ii) Short title

Consonant category in infants

(iii - iv) Authors' names and affiliations

Anne Vilain[1a], Marjorie Dole[1a], Hélène Lœvenbruck[2], Olivier Pascalis[2], Jean-Luc Schwartz[1]

(1) Univ. Grenoble Alpes, Grenoble INP, CNRS, GIPSA-lab, Speech & Cognition Department, 38000 Grenoble, France

(2) Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

Corresponding author

Jean-Luc Schwartz Gipsa-Lab, Département Parole et Cognition Domaine Universitaire

BP46 38402 Saint Martin d'Hères cedex Tel : +33476824128

e-mail : jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

(v) Conflict of Interest Statement

The authors declare no conflict of interest

(vi) Acknowledgements

The research leading to these results has received funding from the European Research Council under the European community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no 339152). The authors would like to thank Christophe Savariaux and Coriandre Vilain for their assistance with the video stimuli, David Méary for help with the protocol design, Marie Sarremejeanne for her technical help and all the parents and the infants who participated in this study. The babies were tested at Babylab Grenoble.

(a) These authors contributed equally to the realization and analysis of the experimental study

The role of production abilities in the perception of consonant category in infants

Research highlights

- **It is still unclear how infants can acquire perceptual categories related to consonant place of articulation in spite of contextual variability**
- **Using an intersensory matching procedure we show that infants around 9 months are able to categorize consonant place of articulation across different vowel contexts**
- **This ability is only present in infants who produce the corresponding consonants in babbling**
- **This shows for the first time that articulatory/motor information provided by babbling helps infants build perceptual speech categories**

ABSTRACT

The influence of motor knowledge on speech perception is well established, but the functional role of the motor system is still poorly understood. The present study explores the hypothesis that speech production abilities may help infants discover phonetic categories in the speech stream in spite of coarticulation effects. To this aim, we examined the influence of babbling abilities on consonant categorization in 6- and 9-month-old infants. Using an intersensory matching procedure, we investigated the infants' capacity to associate auditory information about a consonant in various vowel contexts with visual information about the same consonant, and to map auditory and visual information onto a common phoneme representation. Moreover, a parental questionnaire evaluated the infants' consonantal repertoire. In a first experiment using /b/-/d/ consonants, we found that infants who displayed babbling abilities and produced the /b/

and/or the /d/ consonant in repetitive sequences were able to correctly perform intersensory matching, while non-babblers were not. In a second experiment using the /v/-/z/ pair, which is as visually contrasted as the /b/-/d/ pair but which is usually not produced at the tested ages, no significant matching was observed, for any group of infants, babbling or not. These results demonstrate, for the first time, that the emergence of babbling could play a role in the extraction of vowel-independent representations for consonant place of articulation. They have important implications for speech perception theories as they highlight the role of sensorimotor interactions in the development of phoneme representations during the first year of life.

KEYWORDS

Perception-production link; infants; phoneme categorization; intersensory matching; consonant place of articulation; babbling

MANUSCRIPT

1. Introduction

Phoneme representation in Auditory, Motor and Perceptuo-Motor theories of speech perception

Understanding the nature of phoneme representations remains an outstanding challenge for speech perception theories. The process by which a listener extracts information from the acoustic signal to identify phoneme categories remains largely unclear, because of the complexity of the mapping between sounds and phonemes. Contrasting theories have been developed to explain the process of phoneme categorization. Auditory theories propose that the acoustic signal is directly matched to phonemic representations. The basic cues underlying phoneme identification would thus be purely auditory, independently from any use of motor information (Diehl, Lotto & Holt, 2004). In contrast, motor theories (Galantucci, Fowler, & Turvey, 2006; Liberman, 1957; Liberman & Mattingly, 1985) argue that phoneme identification proceeds through a systematic recoding of the sensory input in terms of the articulatory gestures that are used to produce the speech sounds.

Neurophysiological studies have brought evidence that motor procedural knowledge is activated during speech perception tasks. Brain areas involved in the planning and execution of speech gestures are recruited during visual, audio-visual, as well as purely auditory speech perception tasks (Möttönen, Järveläinen, Sams, & Hari, 2004; Ojanen et al., 2005; Pulvermüller et al., 2006; Wilson, Saygin, Sereno, & Iacoboni, 2004; see Skipper, Devlin & Lametti, 2017 for a recent review). The view that perception and production systems are closely linked is also strengthened by the finding that perturbing the motor system before or during a perception task can modify the perceptual decision

(d'Ausilio et al., 2009; Ito, Tiede, & Ostry, 2009; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Möttönen, Dutton, & Watkins, 2013; Sato et al., 2011). However, none of these data provide a clear answer to the question of the putative functional role of the motor system in the speech perception process.

In this context, integrative perceptuo-motor theories have been introduced, suggesting that speech perception could rely on both auditory and motor mechanisms (Skipper, Van Wassenhove, Nusbaum, & Small, 2007; Schwartz, Basirat, Ménard, & Sato, 2012). Inspired by the classical model of speech recognition by Stevens & Halle (1967), that incorporated an active procedure of "analysis-by-synthesis", Skipper et al. (2007) claimed that the functional role of the motor system in speech perception is to constrain the ultimate phonetic interpretation. According to Skipper et al. (2007), motor system activity constitutes a hypothesis about the phonemes produced, and this hypothesis predicts the sensory consequences of executing that hypothesis through efference copy. These sensory consequences can then be matched with incoming sensory speech input to constrain interpretation. Going one step further, Schwartz et al. (2012) introduced the "Perception-for-Action-Control Theory" which assumes that speech units are intrinsically sensory-motor and result from a co-structuration of the motor and auditory knowledge acquired in the course of language development. They proposed that during language development, motor experience would progressively be combined with the early multi-sensory abilities available at birth, structuring and enriching phonetic representations that would hence include auditory, visual, somatosensory as well as motor features. The present study attempts to test this hypothesis by evaluating the relationship between speech motor abilities and the perceptual categorization of phonetic units in the first stages of phonetic development.

The development of speech perception in relation with speech production abilities

A slew of studies have established that as early as one month of age, infants are able to discriminate speech sounds on the basis of phonetic cues that correspond to adult categories (Eimas & Miller, 1980; Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Jusczyk, Copan, & Thompson 1978; Jusczyk & Thompson, 1978). Since infants at that age do not possess the ability to control their vocal tract for speech production, these results have been considered as evidence that speech perception develops independently of motor experience. This was used as a possible argument against motor theories of speech perception, and led the authors of the Motor Theory to postulate an innate link between speech perception and speech production (Liberman & Mattingly, 1985).

Yet, more recently, a number of studies have examined whether articulatory/motor experience could play a role in the development of speech perception. DePaolis, Vihman, & Keren-Portnoy (2011) investigated the relationship between perception and production abilities around the onset of canonical babbling (by 7 months) in English-acquiring infants. They showed that the number and type of consonants mastered by infants affects their attentional response to speech input. These results on the link between speech babble abilities and speech processing were replicated in Italian infants (Majorano, Vihman, & DePaolis, 2014). A recent study by Hoareau, Yeung & Nazzi (2019) shows that 8-month-old infants with greater production abilities are more efficient in statistical word segmentation.

Another important argument in favor of a functional role of the perception-production link in language development has been brought by research on audio-visual perception. There is a large amount of data showing successful audio-visual matching in infants (Kuhl & Meltzoff, 1982,

1984; MacKain, Studdert- Kennedy, Spieker, & Stern, 1983; Patterson & Werker, 1999, 2002; Yeung & Werker, 2013), suggesting that infants can recognize the correspondence between auditory and visual articulatory stimuli. It has been suggested that the mapping between audio and visual speech stimuli uses common articulatory representation (Kuhl & Meltzoff, 1984). This is in line with Yeung & Werker's (2013) results that showed that when 4-month-old infants were chewing on an object inducing spreading vs rounding lip movements, their audio-visual matching of vowels varied.

The finding that the production of orofacial movements influences infants' perception is strengthened by a recent study (Bruderer, Danielson, Kandhadai, & Werker, 2015) in which 6-month-olds were presented with an auditory non-native Hindi contrast /da/ - /ɖa/ that differs in tongue tip placement (retroflex vs. non-retroflex), while having a teether in their mouth or not. Infants who were given a tongue-tip-constraining teether were not able to discriminate the contrast. This suggests that auditory discrimination of a non-native contrast is impaired when the tongue movement necessary to produce it is prevented. Taken together, these studies strongly suggest that sensorimotor information influences the way infants perceive phoneme information and that the speech production system shapes speech perception early in life.

Neurophysiological studies provide complementary data on the relations between perceptual and motor processes in speech development. Auditory-articulatory cortical connections are present early in life, before any speech motor activity and hence before any possibility of perceptuo-motor learning (Dehaene-Lambertz et al., 2006; Mahmoudzadeh et al., 2013; Perani et al., 2011). Still, these connections seem to strengthen from 6 to 12 months of age (Imada et al., 2006; Perani et al., 2011), when most infants begin to produce adult-like vocalizations. This indicates a reinforcement

of the connections between motor and auditory brain areas related to the development of verbal production and perception. Magneto-encephalography data on the brain activity of infants exposed to native vs. non-native speech reveal an evolution in the role of frontal motor areas in speech perception from 7 to 11 months of age (Kuhl, Ramirez, Bosseler, Lin & Imada, 2014). It is only at 11-12 months that infants' frontal areas are involved in processing non-native speech stimuli, as they are in adults. This suggests a maturation of auditory-articulatory connections, with a potential tuning by motor development (after the onset of babbling around 7 months).

Phoneme categorization as an emergent perceptuo-motor process in speech development

Vowel categories are well defined in acoustic terms (see e.g. Schwartz et al., 1997) and they can be acquired from early auditory representations available at birth (if not before). Indeed, vowel categories emerge early in perceptual development (see e.g. Grieser & Kuhl, 1989) and evidence for perceptual narrowing for vowels is found as early as 6 months of age (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992).

Consonant categories, and typically plosives, however, are less easy to describe acoustically. Several hypotheses have been advanced to define plosive places of articulation in acoustic terms (e.g. by Stevens and colleagues: see Blumstein & Stevens, 1979; Stevens, 1980; Stevens & Blumstein, 1978; or by Sussman and colleagues: see Sussman, McCaffrey, & Matthews, 1991; Sussman, Hoemeke, & Ahmed, 1993; Sussman, Fruchter, Hilbert, & Sirosh, 1998), none of which is really conclusive, in the sense that no acoustic theory or model can adequately explain how a naive listener could categorize the plosive acoustic space in a way compatible with natural places of articulation.

The difference between the acoustic properties of vowels and plosives is explained in Figure 1 (adapted from Laurent, Barnaud, Schwartz, Bessière & Diard, 2017, Fig. 10), which provides the typical representation of vowels in terms of acoustic F1-F2 formants (Figure 1, top) and of plosives in vowel (V) contexts, in terms of F2-F3 formants (Figure 1, bottom). As shown in Figure 1 (top), the pattern for vowels is rather simple. Within the set of all possible (F1, F2) pairs constituting the "articulatorily attainable space" in grey, vowels [i a u] constitute three natural classes, that are easy to separate and categorize. However, the pattern is quite different for plosives. Although the nine plosive-vowel sequences [ba bi bu da di du ga gi gu] correspond to nine distinct items, there is no easy way to group them into 3 classes that correspond to the natural articulatory classes "bilabial" /b/, "coronal" /d/ and "velar" /g/ (Figure 1 bottom).

The claim in the Perception-for-Action-Control Theory (Schwartz et al., 2012; Laurent et al., 2017) is that when infants begin to babble, at around 7 months, they discover the articulatory gestures and configurations associated with [bV], [dV] and [gV] syllables and they realize that these configurations belong to three natural *articulatory* – though *not acoustic* – classes. Hence the proposal that phonetic classes are defined by mixed acoustic/auditory, somatosensory, and articulatory/motor properties that are learnt during the joint development of speech perception and production. A prediction following this proposal is that while vowels can be learnt early in development on the basis of purely acoustic configurations, plosives cannot emerge as a set of phonetic classes organized by place of articulation (e.g. /b/, /d/, /g/) until infants begin to discover these places of articulation in their own productions, basically after the onset of babbling at 7 months of age.

The question of the age at which infants become able to categorize consonants, when the vowel context is varied, is still an open one (Jusczyk & Derrah, 1987). In fact, a number of studies have failed to show convincing evidence that infants could have access to invariant phonemic representations for plosives independent of vowel context in the first months of life (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Eimas, 1999). Two recent studies, however, claimed to have provided data on plosive categorization in infants before the onset of babbling. Firstly, Hochmann & Papeo (2014) provided data on plosive categorization in 6-month-old infants. They recorded pupil dilation while presenting sequences of three monosyllabic words that began with the same consonant [b] or [d] followed by three different vowels (e.g. with [d]: 'deed', 'dad', 'dote'), and then a fourth word that, in standard conditions, began with the same consonant followed by a fourth vowel (here [d] in 'due'). In the deviant condition, the fourth word started with the other consonant, in the same fourth vowel context (here [b] in 'boo'). They found that 6-month-old infants' pupil diameter changed in the deviant condition, suggesting that as early as 6 months of age, infants were able to form a category for the onset consonant in spite of the varying vowel context and of the consequential lack of acoustic invariance in the consonant. Since only three of the fourteen 6-month-old infants who participated in the study had entered the canonical babbling phase, the authors concluded against a strong version of the motor theory, according to which the invariance problem is solved through the evocation of motor representation. Some infants could indeed perceive the common onset consonant in syllables that they had never produced.

However, the problem with Hochmann & Papeo's (2014) paradigm is that it might not actually deal with *categorization* but with *discrimination*. To make this clear, let us consider one of the two test

sessions in their experiment, which consists in comparing reactions to a reference series [di da do du] to reactions to a test series [di da do bu]. As displayed in Figure 1, alveolar configurations [di da do du] are located in the left region of the (F2, F3) space with larger F2 and F3 values ([do] is not represented in the figure, but it is located between [du] and [da]). On the contrary, the test stimulus [bu] is the configuration with the lowest F2 and F3 values at the bottom-right corner. Therefore, the test series [di da do bu] displays indeed more acoustic variance than the reference series [di da do du], and acoustic discrimination alone is sufficient to explain their results.

Another study recorded high-density event-related potentials to examine infant categorical perception (Mersad & Dehaene-Lambertz, 2015). The results show that 3-month-old infants presented with CV syllables consisting of a stop consonant [b] or [g] followed by a vowel within the set [a ɛ ɑ̃ ɛ̃] presented larger mismatch responses to a syllable with a new vowel [i] if the syllable also involved a change in consonant (from [b] to [g] or from [g] to [b]). The authors concluded that infants at 3 months, before the onset of babbling, "can compute automatically consonant representation, independently of the vocalic context". Still, the same argument can be raised against this interpretation, as local proximities could well explain these data rather than the hypothesis that invariant representations of stop consonants are computed by infants. Here again, it is likely that [gi] is closer to [ga gɛ gɑ̃ gɛ̃] than [bi] is, and vice versa.

In sum, both the studies by Hochmann & Papeo (2014) and by Mersad & Dehaene-Lambertz (2015) merely deal with acoustic distances between a context set and a test item. Testing genuine categorization in infants requires a paradigm in which they would have to group together items which are spatially dispersed in the acoustic space, as in Fig. 1. This is the objective of the experimental paradigm introduced in the present study.

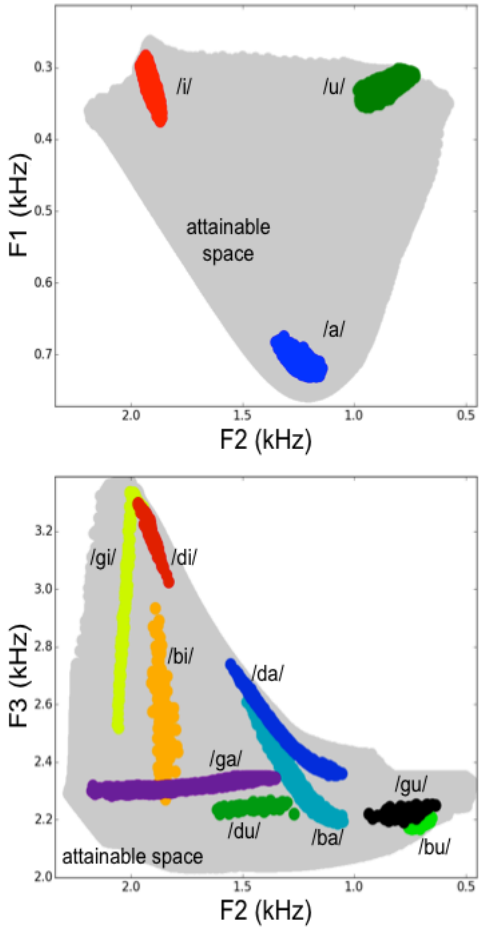

Figure 1 – ***Acoustic representations of vowels and plosives***

The (F1, F2) (top) and (F2, F3) (bottom) articulatorily attainable space, in grey, with dispersion regions for vowels (top) and consonants in CV contexts (bottom), with C within [b d g] and V within [i a u] (from Laurent et al., 2017, Fig. 10).

An original paradigm based on intersensory matching

Our paradigm is based on the intersensory matching procedure developed by Pons, Lewkowicz, Soto-Faraco, and Sebastián-Gallés (2009) who investigated the perceptual narrowing of intersensory matching abilities during the first year of life. Their procedure involved presenting

two side-by-side videos where a speaker silently and repetitively uttered a /ba/ or a /va/ syllable after an auditory familiarization with one of the two syllables. The authors found that by the age of 6 months infants were able to successfully perform intersensory matching and looked longer at the video matching the previously presented auditory consonant. In the present study, we adapted this procedure in order to investigate the development of consonant categorization: instead of presenting the same syllable in the audio familiarization and in the video test, we presented a syllable with the same consonant, but with a different vowel. We thus evaluated the ability of 6- and 9- month-olds to match a series of audio syllables such as [bi be bo…] with constant consonant and varying vowel, with a visually presented syllable such as [ba] vs. [da], i.e. with a novel vowel environment (Experiment 1). This intersensory category matching procedure tests the ability to identify the common consonant in the audio stimuli, and to relate it with the consonant in the visual stimulus, and to do this while factoring out the contextual effects produced by the following vowel. It rules out the possibility of categorization being performed by mere detection of acoustic cues, and it provides a way to assess the emergence of multisensory/motor phoneme representations.

Since our assumption is that motor experience is required to elaborate an adequate representation of plosive place of articulation, we predicted that only infants who had started producing the adequate consonants in the course of babbling would succeed in the task. Babbling being a stage in speech production development when infants start producing repeated proto-consonant-vowel sequences with high frequency (MacNeilage & Davis, 2001), we hypothesized that it constitutes a critical period for motor experience and auditory-motor relationship. To test this hypothesis, we documented the babbling abilities of our participants by means of a questionnaire addressed to parents.

Finally, to rule out the possibility that the infants' performance may merely be based on categories derived from visual experience with no involvement of their own speech production system, we replicated the intersensory matching procedure using another contrast, /v/-/z/ (Experiment 2). Indeed, this contrast is also highly visible but typically not produced by infants between 6 and 9 months of age, contrary to /b/ and /d/. We predicted that if the motor system is indeed crucial to the development of phoneme representation, infants should not be able to extract phonemic cues with this second contrast in spite of its high visibility.

2. Methods and Results

2.1. Experiment 1

2.1.1. Participants

Twenty 6-month-olds (9 females) (M_{age}=191.8 days, SD_{age}=4.5 days) and 25 9-month-olds (15 females) (M_{age}=285.6 days, SD_{age}=4.6 days) were included in the analyses. All participants were full-term infants, recruited from the maternity hall of the Centre Hospitalier Universitaire, Grenoble, Alpes, France. All lived in a French-speaking environment. Thirty-two additional infants were tested, but were excluded from analyses because they heard less than 90% French at home (4 infants), or due to fussiness (12 infants), or because their parents had failed to fill in the report on speech production (16).

2.1.2. Stimuli

Video stimuli

Stimuli were composed of two side-by-side video recordings of the same native speaker of French silently uttering /ba/ and /da/ at a rate of 1 syllable per second. To ensure that idiosyncratic characteristics of the talker's production would not influence the infants' behavior,

two different female speakers, aged 25 and 26, were recorded. The children were divided into two groups: one group of children only saw the first speaker while the second group only saw the second speaker. Video recordings were sampled at a 50 Hz sampling rate. During recording, the speakers were asked to directly look at the camera and to keep a neutral expression while repeating the syllables at a comfortable rate. After recording, the videos were edited, and one exemplar of each syllable was selected and looped every 1s in order to obtain sequences lasting 21s. The two videos corresponding to two different syllables were pasted side by side to create a stimulus for the baseline and test trials. During editing, we ensured that the two videos started with the same mouth configuration, in order to obtain a correct synchronization between both facial movements. Faces were recorded against a blue background. The final size of each video was 18 cm wide and 20 cm high.

2.1.3. Audio stimuli

In order to avoid any bias due to idiosyncratic cues in the talker's production, auditory stimuli were composed of recordings of 5 new female speakers aged 20 to 32, repeatedly pronouncing syllables containing /d/ or /b/ consonants associated with the 4 vowels /i/, /e/, /u/ and /o/ (no /a/ in the auditory stimuli). Two repetitions per speaker were recorded, then randomly mixed and concatenated to form 42s-long multi-speaker sequences; syllables were presented at a rate of 1 syllable per second. Stimuli were digitally recorded using a PMD Marantz recorder with a high-quality audio microphone in a soundproof room at a 44.1 kHz sampling rate and normalized at a 70 dB intensity level.

2.1.4. Procedure

Infants were seated on their parents' laps in a dimly illuminated room, 60 cm away from a 22-inch computer screen. All parents signed a written consent form prior to the experiment and infants received a book for their participation. The study was evaluated by the Ethics Committee of Université Grenoble Alpes (CERNI) and received a positive evaluation (avis 2014-03-11-38). The parents were asked not to intervene or interact with the infant during the entire experiment, and they were unaware of the objective of the experiment.

The experiment consisted in six trials (see Figure 2). In order to take into account potential baseline preference for any of the two silent videos, the experiment started with two Baseline trials during which the two side-by-side silent videos of the same speaker (one for /ba/, one for /da/) were presented during 21s (trials 1 and 2). These Baseline trials were followed with two 42s auditory familiarizations (trials 3 and 5) during which infants heard several speakers uttering one of the two consonants (/b/ or /d/), associated with different vowels. For each age, infants were divided into two groups, one presented with only /b/ (10 6-month-olds, 13 9-month-olds) and the other with only /d/ (11 6-month-olds, 12 9-month-olds).

During the auditory presentation, an attention getter consisting of a moving ball with changing color and size was presented to the infants in order to keep their attention to the screen. Two test trials (trials 4 and 6) followed, during which infants were presented with the two same side-by-side videos as in the Baseline. Between trials 1 and 2, and trials 4 and 6, the side of the syllable presentation (/ba/ or /da/) was reversed, and the order was counterbalanced between participants.

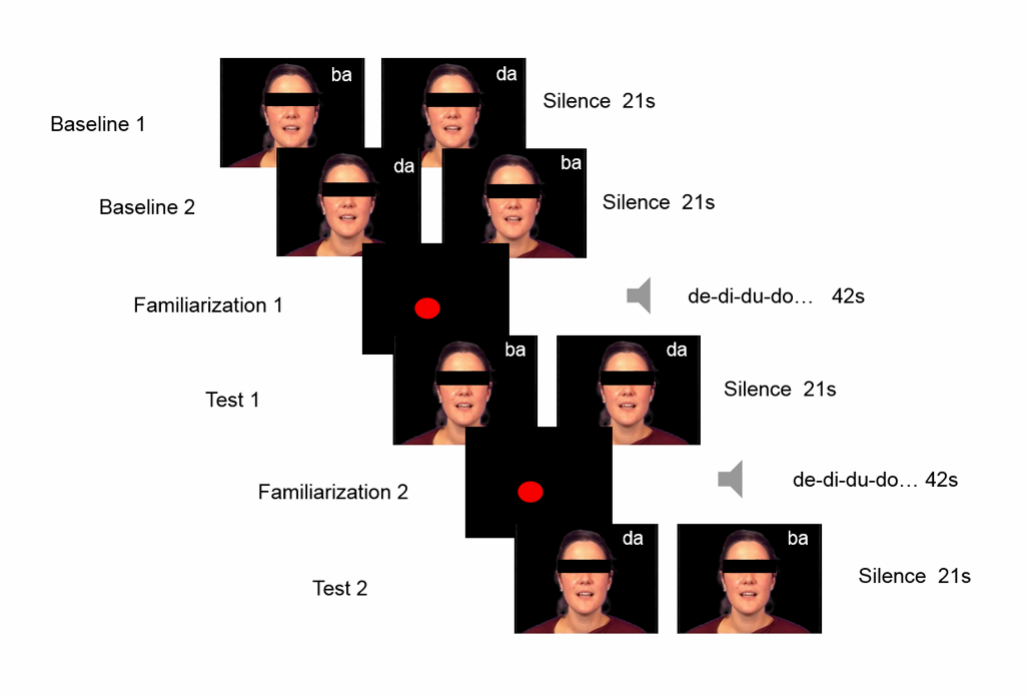

Figure 2. Schematic representation of the intersensory matching procedure*. Only one auditory condition is shown (here, familiarization with the consonant /d/).*

The experiment was run using the E-prime 2.0 software (Psychology Software Tool, Pittsburgh, PA, USA). Two loudspeakers (Dell A225) were placed behind the screen, to play the auditory stimuli. We used a low-light video camera, located above the screen to record the infants' looking behavior. Video recordings were then digitized and analyzed using a frame-by-frame coding procedure. The mean looking time on each video and for each trial was measured for each infant.

2.1.5. Parental Questionnaire

Parents were also asked to fill in a questionnaire assessing the vocal productions of their infant. The questionnaire was sent to the parents at least one week before testing. It consisted of a list of 10 consonants (/b/, /d/, /g/, /p/, /t/, /k/, /l/, /m/, /n/, /s/). Parents had to evaluate whether their infant produced each consonant, and if so, in what type of syllable sequence they produced it (*one*

syllable/two syllables/more than two syllables), to judge the frequency of production (*never/sometimes/often/very frequently*), and to describe the vowels associated with this consonant *(/a/, /i/, /o/, /u/, /e/, or other*). We computed two different babbling scores for each infant, one characterizing the production of babbling in general, that is with any of the 10 reported consonants (Babbling_general), and the other one narrowing to the two consonants in the categorization test, namely /b/ and /d/ (Babbling_bd). The infants were assigned to one of two groups according to their production abilities on the /b/ - /d/ contrast. For the *Babbling_general* score, the infants who did not produce any repetitive sequence (two syllables and more) were included in the Non-Babbling group. The infants who produced repetitive sequences, with any of the consonants in the questionnaire, were included in the Babbling group. For the *Babbling_bd* score, the infants who produced none of the /b,d/ consonants in repetitive sequences (two syllables and more) were included in the Non-Babbling group. The infants who produced at least /b/ or /d/ in repetitive sequences were included in the Babbling group.

For the Babbling_general score, the Non-Babbling group was composed of 13 infants (thirteen 6-month-olds and no 9-month-olds), and the Babbling group was composed of 32 infants (seven 6-month-olds and twenty-five 9-month-olds). For the Babbling_bd score, the Non-Babbling group was composed of 17 infants (fifteen 6-month-olds and two 9-month-olds), and the Babbling group was composed of 28 infants (five 6-month-olds and twenty-three 9-month-olds).

2.1.6. Data pre-processing

For each individual infant and each trial, the looking time (LT) towards the "ba" and "da" videos was recorded. A proportion of LT was computed for each video, as the ratio in percent between looking time towards that video and the total looking time to both videos. For example for the /ba/ syllable, we computed: $\%LT_{ba}=LT_{ba}/(LT_{ba}+LT_{da})*100$. A Difference Score for the matching face was calculated between the proportion of LT in the two Test trials and in the two Baseline trials (proportion of total time that infants spent looking at the matching face during the two Test trials minus proportion of total time that they spent looking at the matching face during the two Baseline trials ="matching score"). The "matching face" was defined according to the category of the audio stimuli presented in the familiarization phases 3 and 5. Thus a positive matching score reflected a preference for the matching face whereas a negative one reflected a preference for the non-matching face. Following the results obtained by Pons et al. (2009), we expected the proportion of LT directed at the matching syllable to be greater during the Test than during the Baseline in infants who made successful intersensory matches.

2.1.7. Analysis and results

A linear model was run, with matching score as the dependent variable, and age (6 vs 9 months), babbling stage (*Babbling_bd* score: non-babbling_bd vs babbling_bd), gender (male vs female), familiarization consonant (consonant heard during auditory familiarization, /b/ or /d/), and their interactions as factors. The model was fitted with the following R code: *lm(MatchingScore~Babbling_bd*Age+Gender+FamiliarizationCons)*. Then a variable selection procedure was applied, which led to retain Babbling_bd as the only significant effect ($F(1,43) = 6.31$, $p= 0.016$, Cohen's $d= 0.73$). Neither age, nor gender or familiarization consonant had significant effects, alone or in interaction, with negative inclusion tests for Babbling_bd*Age ($p=0.781$), Age ($p=0.070$), Gender ($=0.539$), and FamiliarizationCons ($p=0.432$). The selected model, lm(MatchingScore~Babbling_bd), was then validated with a residual analysis. Finally, one-tailed t-tests against zero were run to test if the matching scores for the two babbling groups were positive. The t-tests revealed that the matching scores for the non-babbling infants were not different from zero ($t (16)= -1.64$, $p = 0.93$), whereas they were positive for the babbling infants ($t(27)=1.91$, $p= 0.033$, Cohen's $d= 0.36$). The matching scores for the two babbling groups are presented in Figure 3. The t test against zero for the 6-month-olds' matching scores was negative ($t(19) = -1.10$, $p = 0.86$), as was the one for the 9-month-olds ($t(24)= 1.56$, $p = 0.066$).

The same analysis was run with the *Babbling_general* score (indicating whether the infant babbled with any type of consonant). This time, neither age, gender, familiarization consonant, nor babbling score was retained as meaningful variable ($F(1,43)= 2.40$, $p= 0.128$) for *Babbling_general*).

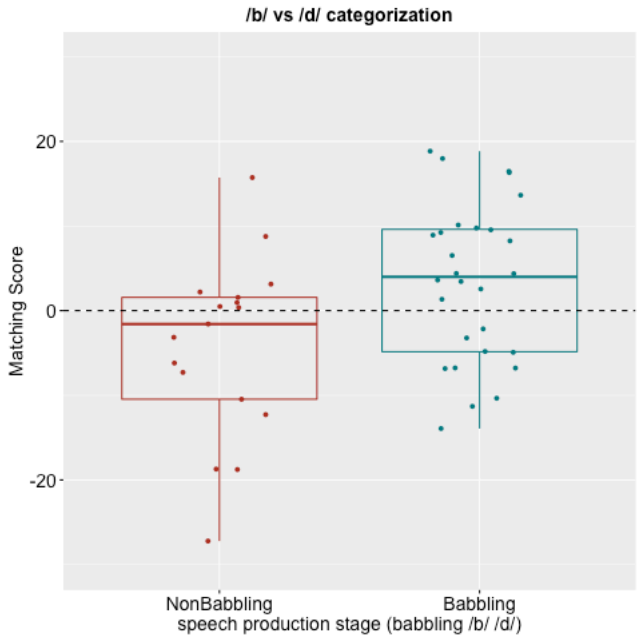

Figure 3. Matching scores for the /b/ - /d/ contrast as a function of production abilities *(Babbling_bd score). Positive matching scores indicate successful intersensory matching. The points represent individual data.*

2.1.8. Discussion of Experiment 1

The aim of the present study was to investigate the relationship between phoneme categorization and babbling abilities. We employed an intersensory matching procedure in order to test infants' ability to link auditory and visual consonantal information and to map it onto a single crossmodal representation.

Our results show that the development of production abilities has a significant effect on phoneme perception. Indeed, infants who already produced at least one of the two consonants of interest in babbling sequences exhibited matching abilities whereas non-babbling infants did not. On the contrary, age had no significant effect on infants' performance, and 9-month-old infants did

not display better matching scores than 6-months-olds. This suggests that the ability to produce the consonants presented here allowed infants to extract the common consonantal information in all stimuli, presumably using a motor representation associated with both the auditory and the visual inputs. This implies that motor information may help defining phonetic representations.

At this stage, however, two alternative interpretations might be suggested. A first possibility is that infants who are more advanced in terms of production would also show better matching scores, due to improved general cognitive and/or linguistic abilities, or more advanced motor control skills, which would not be related to the perception-production link. The hypothesis that general motor control skills alone can explain the data is contradicted by the fact that, in this study, the general babbling abilities (*Babbling_general* score) of the infants were unrelated to their categorization performance. This indicates that higher matching scores cannot simply be explained by an improvement in general motor abilities. A second possibility is that the intersensory matching would only be based on the processing of audio-visual co-occurrences, without resorting to motor processes. When presented with audio /bV/ sounds, older infants could successfully recover visually opening/closing lips, and with /dV/ sounds they could simply evoke a visual sequence consisting of jaw opening/closing cycles with a visible tongue tip. Intersensory matching could simply rely on the learning of audiovisual associations – which infants are capable of, as mentioned above – without the requirement of motor knowledge acquired over the course of speech production development. Therefore, in order to control for these two possible confounding factors, we ran a second experiment with the /v/-/z/ contrast instead of the /b/-/d/ contrast (Experiment 2). The /v/-/z/ contrast was chosen for four reasons. First, it is visually very close to the /b/ - /d/ contrast tested in our first experiment, which allowed us to control for a purely audio-visual effect (where jaw-

lip gestures in /b/ and /v/ would be contrasted with jaw-tongue tip gestures in /d/ and /z/). Second, it is visually as salient as the /b/-/d/ contrast, as proven by studies on consonantal visemes (visually distinct categories) in various languages. Descriptions of e.g. American English (Binnie, Montgomery & Jackson, 1974; Fisher, 1968; Walden, Prosek, Montgomery, Scherr & Jones, 1977), Dutch (van Son, Huiskamp, Bosman & Smoorenburg, 1994) or French (Gentil, 1981; Benoît, Mohamadi & Kandel, 1994) all demonstrate that /b/, /v/, /d/ and /z/ belong to four different visual categories, and that these four stimuli belong to different main branches in the visual confusion tree published by Summerfield in his inspiring review of visual speech perception (Summerfield, 1987). Third, although fricatives have sometimes been considered less easily discriminated than plosives in early infancy, a convergent bundle of studies show that infants can discriminate fricative place of articulation at an early age, and particularly that they are well able to discriminate labiodental from coronal fricatives before 6 months of age (see e.g. Eilers & Minnifie, 1975; Holmberg, Morgan & Kuhl, 1977; Levitt, Jusczyk, Murray, & Carden, 1988; Beach & Kitamura, 2011). Fourth, and crucially, contrary to /b/ and /d/ that are among the first consonants to appear in infants' inventories, /v/ and /z/ appear very late in the development of speech production and it is quite unlikely to find these consonants in the babbling stage (e.g. Locke, 1983; Kern, Davis & Zink, 2009). Therefore, our prediction in Experiment 2 is that since infants in the 6-9 months period presumably do not produce articulatory configurations typical of /v/ and /z/, they should not display intersensory category matching with these consonants, independently of their babbling ability.

2.2. Experiment 2

2.2.1. Participants

Twenty-five 6-month-olds (14 females) (M_{age}=193.7 days, SD_{age}=6.8 days) and 25 9-month-old infants (16 females) (M_{age}=286.1 days, SD_{age}=5.9 days) participated in this study. Thirty-one additional infants were tested but not included in the final analyses due to fussiness (6 infants), or because they heard less than 90% French at home (1 infant), or because their parents had failed to fill in the report on speech production (24).

2.2.2. Stimuli

Video stimuli

Stimuli were composed of two side-by-side video recordings of the same native speaker of French silently uttering /va/ and /za/ at a rate of 1 syllable per second. Two female speakers, aged 26 and 35 and different from those of Experiment 1 were recorded. The procedure used to build the stimuli was the same as in Experiment 1.

Audio stimuli

Auditory stimuli were composed of recordings of 3 speakers aged 24 to 35, repeatedly pronouncing syllables with /v/ or /z/ at the onset, associated with 4 different vowels, /i/, /e/, /u/, /o/. Four repetitions per speaker were recorded, randomly mixed and concatenated to form 42s-long multi-speaker sequences; syllables were presented at a rate of 1 syllable per second. Stimuli were recorded using a PMD Marantz recorder with a high-quality audio microphone in a soundproof room at a sampling rate of 44.1 kHz and normalized at a 70dB intensity level.

2.2.3. Procedure

The procedure was identical to that in Experiment 1. The experiment consisted in six trials: 2 Baseline trials with two side-by-side silent videos of one speaker repeatedly pronouncing /va/

and /za/ (trials 1 and 2); 2 auditory familiarizations (trials 3 and 5) lasting 42s during which infants were auditorily familiarized with CV syllables containing one of the two consonants (/v/ or /z/), followed by 4 different vowels /i/, /e/, /u/, /o/ (randomized orders); and 2 test trials (trials 4 and 6) during which infants were presented with the same two side-by-side videos as in the Baseline. Between trials 1 and 2, and trial 4 and 6, the side of syllable presentation was reversed. Once again, infants for each age were separated into two groups, half infants heard the /v/ consonant (11 6-month-olds, 13 9-month-olds), whereas the other half heard the /z/ consonant (14 6-month-olds, 12 9-month-olds).

2.2.4. Parental Questionnaire

As in Experiment 1, parents were asked to fill in a questionnaire assessing the production abilities of their infants. A list of 12 consonants (/b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /f/, /v/, /s/, /z/) was presented. As in the previous experiment, parents were asked to note if their infant produced each consonant, and the type of syllable sequence they produced it in, to evaluate the frequency and to describe the vowels associated with this consonant.

On the basis of these production questionnaires, we characterized the infants with regards to their production of the /b/ and or /d/ consonants, as we had done in the first experiment, since this second experiment was run as a control for the previous one. Experiment 1 suggested that infants who had /b/ and /d/ in their babbling repertoire were better at categorizing these consonants because they had gained articulatory knowledge of these consonants, and not because they had better general production abilities than the others. Experiment 2 was designed to further test this hypothesis, by showing that infants who master the production of /b/ and /d/ sequences but do not yet produce fricatives cannot categorize fricative pairs of

consonants, such as /v/ vs /z/, for which they have no articulatory experience. The infants were assigned to one of the following two groups, according to their production abilities on the /b/ - /d/ contrast (*Babbling_bd* score): i) *Non-Babbling*: infants who produced neither the /b/ nor the /d/ consonant in repetitive sequences; ii) *Babbling*: infants who produced the /b/ and/or the /d/ consonant in repetitive sequences (two syllables and more). The *Non-Babbling* group was composed of 19 infants (18 6-month-olds and 1 9-month-olds), the *Babbling* group was composed of 31 infants (7 6-month-olds and 24 9-month-olds).

Importantly, we also checked the ability of the two groups of infants to produce the /v/ and/or the /z/ consonant in repetitive sequences. As was expected, only a very small number of infants display this ability, respectively 0 in the Non-Babbling group and 6 in the Babbling group (one 6-month-olds, and five 9-month-olds). Therefore, the prediction at this stage is that if general linguistic/cognitive maturity properties predict the intersensory matching ability, the same results should be obtained in Experiment 2 as in Experiment 1. On the contrary, if the ability to produce at least one of the two involved consonants is required to perform the task, no intersensory matching ability should be observed for either group in Experiment 2.

2.2.5. Analyses and results

As in the first experiment, the *matching score* was calculated between the proportion of Looking Time in the two Test trials and in the two Baseline trials (proportion of total time that infants spent looking at the matching face during the two Test trials minus proportion of total time that they spent looking at the matching face during the two Baseline trials). A linear model was run, with matching score as the dependent variable, and age (6 vs 9 months), babbling stage (non-babbling vs babbling), gender (male vs female), familiarization consonant (consonant heard

during auditory familiarization, /v/ or /z/), and their interactions as factors. The model was fitted with the following R code:

*lm(MatchingScore~Babbling_bd*Age+Gender+FamiliarizationCons)*. Then a variable selection procedure was applied, which led to retain no significant factor. Neither babbling, nor age, gender or familiarization consonant had significant effects, alone or in interaction: the inclusion tests were negative for *Babbling_bd*Age* (p=0.486), *Age* (p=0.727), *Babbling_bd* (p=0.820), *Gender* (=0.662), and *FamiliarizationCons* (p=0.053). To illustrate the data, the matching scores for the two babbling groups are presented in Figure 4. Finally, one-tailed t-tests against zero were run to test if the matching scores for the two babbling groups were positive. The t-tests revealed that the matching scores were not different from zero, either for the non-babbling infants ($t(18) = 0.41$, $p = 0.343$) or the babbling infants ($t(30) = 0.22$, $p = 0.414$). The t test against zero for the 6-month-olds' matching scores was negative ($t(24) = 0.5$, $p = 0.31$), as was the one for the 9-month-olds ($t(24)= 0.07$, $p = 0.472$).

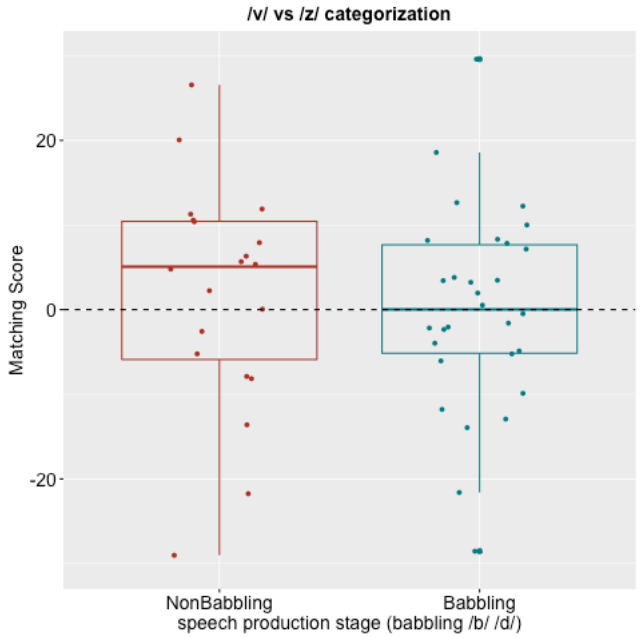

Figure 4. Matching scores for the /v/ - /z/ contrast as a function of production abilities *(non-babbling vs babbling). The points represent individual data.*

In sum, infants at 6 or 9 months of age seem to be unable to categorize /v/ or /z/ consonants, which are largely absent from their productive inventories. They are unable to do so, even if they have started producing babbling sequences with other types of consonants, which shows that the ability to form phonetic categories develops in relation with production abilities in linguistic development. Indeed, these results support the idea that the ability to categorize consonants builds on articulatory experience with these specific consonants.

3. General Discussion

In this article we investigated the speech perception-production relationship during infancy. More particularly we examined the influence of the development of speech production abilities on phoneme categorization around the onset of babbling. To this aim, we used an intersensory matching procedure to evaluate infants' ability to form a representation for a common consonant in various vowel contexts, and to match it to the visual gesture that is used to produce this consonant. Our results show that only infants who have started babbling and who produce /b/ and/or /d/ in repetitive sequences can perform this matching on /b/ or /d/ consonants. We therefore suggest that the development of production abilities may help infants to refine their perceptual categories and to build reliable phonemic representations. This interpretation is reinforced by the results of our second experiment using a /v/ - /z/ contrast. We used this contrast because it involves consonants that are not generally produced at these ages. If the observed effect of production abilities on the /b/ - /d/ contrast resulted indeed from an involvement of the perception-production loop, and not from improved general cognitive abilities in babbling infants, we expected to find no significant matching for this second contrast. In agreement with these expectations, we did not obtain any significant preference for any of the /v/ or /z/ videos, either for the babbling or for the non-babbling infants. This finding further argues for the hypothesis that when infants start producing a sound, their representation for that sound becomes richer, involving auditory as well as motor information, and as a consequence they display better categorization abilities.

These data converge with a bundle of experimental facts suggesting that the orofacial system does play a role in speech perception by infants in the course of their first year of age. For instance,

production abilities have been shown to play a role in infant's attention to specific aspects of the acoustic input (DePaolis et al., 2011), in the way the content of the acoustic material is processed (Bruderer et al., 2015) or in the ability to match the audio and video contents of a speech audiovisual material (Yeung & Werker, 2013).

Our findings enable us to draw a possible developmental scenario, capitalizing on previous data and proposals. As classically known, infants at birth are able to discriminate speech sounds in a nonlinear way, that is, with better discrimination at specific positions along acoustic continua. This behavior is compatible with the nonlinear discrimination patterns associated with categorical perception data in adults (e.g. Eimas et al., 1971). This ability, observed long before vocalizing and babbling, exploits basic auditory capacities known to be mature at birth. Importantly, such data are associated with patterns of discrimination, and not with a categorization process per se. This is also the case for studies on older infants, showing perceptual narrowing, e.g. for vowels (Kuhl et al., 1992) or consonants (Werker & Tees, 1984), which are all based on discrimination paradigms.

Then, vocalizations in the first months of age would allow infants to acquire some knowledge of the relationships between sounds and orofacial gestures. The orofacial exploration that is known to take place before babbling onset (Kuhl & Meltzoff, 1996), would help infants to develop/refine their ability to discriminate sounds (Bruderer et al., 2015) and to match sounds and sights of a speaker (Yeung & Werker, 2013). These early perceptuo-motor relationships have been shown to be underpinned by a dedicated cortical circuit as early as 2 months of age (Dehaene-Lambertz et al., 2006).

As suggested by the present data, babbling onset around 7-8 months would be the stage at which infants start exploring perceptuo-motor relationships more systematically, with repeated

production of consonant-vowel sequences, and increasing use of variegation. This enriched perceptuo-motor information would help the infants progressively build representations of phonetic units and elaborate perceptuo-motor *categories* – associating auditory cues, e.g. for vowels or voicing, and articulatory/motor cues, e.g. for plosive place of articulation. They would also gradually be used to process speech in adverse conditions (Kuhl et al., 2014).

Finally, since the present data suggest that orofacial knowledge plays a role in the elaboration of phonetic categories, a remaining question concerns the exact nature of this knowledge, and the way it is combined with auditory information to form cognitive representations of speech units. Two different hypotheses can be advanced. First, the orofacial knowledge related to place of articulation in babbling infants in Experiment 1 could be related to motor knowledge: infants would discover that some specific articulatory gestures, i.e. lip closure for /b/ and tongue tip upward movement for /d/, are at play. Therefore, motor knowledge on how sounds are articulatorily produced would be integrated with auditory information, leading to an auditory-motor representation. Alternatively, it can be speculated that the somatosensory system could provide tactile and proprioceptive information on the place of occlusion in the vocal tract, with lip contact for labials or anterior tongue-palate contact for alveolars. This somatosensory feedback would be integrated with auditory information to form a multisensory representation. The present data do not allow us to disentangle these two hypotheses. They actually refer to a longstanding debate in speech science on the distinction between motor gestures and their somatosensory and auditory consequences and on the relative contributions of both in speech perception (e.g. Liberman & Mattingly, 1985; vs. Fowler & Rosenblum, 1991).

Both views are in line with the theoretical framework elaborated in the Perception-for-Action-Control Theory (Schwartz et al., 2012), which argues for the multisensory-motor nature of speech units. In this framework, it is claimed that cognitive representations of speech units combine auditory, visual, somatosensory and motor representations within a bundle of heterogenous features. This view adequately construes the phonology-phonetic interface in a way compatible with neurocognitive data on speech communication (Loevenbruck et al., 2018; Laurent et al., 2017; Patri, Perrier, Schwartz & Diard, 2018). The present data offer experimental evidence in favor of such a multisensory-motor scenario in the course of speech development.

4. Acknowledgements

The research leading to these results has received funding from the European Research Council under the European community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no 339152). The authors would like to thank Christophe Savariaux and Coriandre Vilain for their assistance with the video stimuli, David Méary for help with the protocol design, Marie Sarremejeanne for her technical help and all the parents and the infants who participated in this study. The babies were tested at Babylab Grenoble.

5. Data availability statement

The data that support the findings of this study are openly available in [PhonCat] at:

[https://osf.io/fzb59/?view_only=3b4e399466f34388aa1eec8c33a31ebe],

reference number [fzb59] (Vilain et al., 2019).

6. List of figure legends

Figure 1 – Acoustic representations of vowels and plosives. *The (F1, F2) (top) and (F2, F3) (bottom) articulatorily attainable space, in grey, with dispersion regions for vowels (top) and consonants in CV contexts (bottom), with C within [b d g] and V within [i a u] (from Laurent et al., 2017, Fig. 10).*

Figure 2. Schematic representation of the intersensory matching procedure*. Only one auditory condition is shown (here, familiarization with the consonant /d/).*

Figure 3. Matching scores for the /b/ - /d/ contrast as a function of production abilities *(Babbling_bd score). Positive matching scores indicate successful intersensory matching. The points represent individual data.*

Figure 4. Matching scores for the /v/ - /z/ contrast as a function of production abilities *(non-babbling vs babbling). The points represent individual data.*

7. References

Bahrick, L.E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13*(3): 99–102. doi: 10.1111/j.0963-7214.2004.00283.x.

Beach, E. F., & Kitamura, C. (2011). Modified spectral tilt affects older, but not younger, infants' native-language fricative discrimination. *Journal of Speech, Language and Hearing Research*, *54*(2), 658-667. doi:10.1044/1092-4388

Benoît, C., Mohamadi, T., & Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French speech in noise. *Journal of Speech and Hearing Research, 37*, 1195–1203.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P.W., Kennedy, L.J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech. *Journal of Experimental Psychology. General, 117*(1): 21-33. doi: 10.1037/0096-3445.117.1.21.

Binnie, C.A., Montgomery, A.A., & Jackson, P.L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research, 17*, 619-630.

Blumstein, S.E., & Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America, 66*, 1001–1017. http://dx.doi.org/10.1121/1.383319

Bruderer, A.G., Danielson, D.K., Kandhadai, P., & Werker, J.F. (2015). Sensorimotor influence on speech perception in infancy. *Proceedings of the National Academy of Science USA, 112*(44): 13531-13536. doi: 10.1073/pnas.1508631112.

d'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology, 19*(5): 381-385. doi: 10.1016/j.cub.2009.01.017.

Dehaene-Lambertz, G., Hertz-Panier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., & Dehaene, S. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Science USA, 103*(38): 14240-14245. doi: 10.1073/pnas.0606302103.

DePaolis, R.A., Vihman, M.M., & Keren-Portnoy, T. (2011). Do production patterns influence the perception of speech in prelinguistic infants? *Infant Behavior & Development, 34*(4): 590-601. doi: 10.1016/j.infbeh.2011.06.005.

Diehl, R.L., Lotto, A.J., & Holt, L. (2004). Speech perception. *Annual Review of Psychology, 55*: 149-179. doi: 10.1146/annurev.psych.55.090902.142028.

Eilers, R.E., & Minifie, F.D. (1975). Fricative discrimination in early infancy. *Journal of Speech and Hearing Research, 18*, 158-167.

Eimas, P.D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America, 105*(3): 1901-1911. doi: 10.1121/1.426726.

Eimas, P.D., & Miller, J.L. (1980). Contextual effects in speech perception. *Science*, *209*(4461): 1140-1141. doi: 10.1126/science.7403875.

Eimas, P.D., Siqueland, E.R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968): 303-306. doi: 10.1126/science.171.3968.303.

Fisher, C.G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11*, 796-804.

Fowler, C. A., & Rosenblum, L. (1991). In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33-59). Hillsdale, NJ: Erlbaum.

Galantucci, B., Fowler, C.A., & Turvey, M.T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3):361-377.

Gentil, M. (1981). Etude de la perception de la parole : Lecture labiale et sosies labiaux (*Speech perception study: lipreading and visemes*). Technical Report, IBM, France.

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25,* 577–588.

Guellaï, B., Streri, A., & Yeung, H.H. (2014). The development of sensorimotor influences in the audiovisual speech domain: some critical questions. Frontiers in Psychology, 5: 812. doi: 10.3389/fpsyg.2014.00812.

Hoareau, M. , Yeung, H. H. and Nazzi, T. (2019), Infants' statistical word segmentation in an artificial language is linked to both parental speech input and reported production abilities. *Developmental Science. Accepted Author Manuscript*. doi:10.1111/desc.12803

Hochmann, J.R., & Papeo, L. (2014). The invariance problem in infancy: a pupillometry study. *Psychological Sci*ence, *25*(11): 2038-2046. doi: 10.1177/0956797614547918.

Holmberg, T.L., Morgan, K.A., & Kuhl, P.K. (1977). Speech perception in early infancy: discrimination of fricative consonants. *Journal of the Acoustical Society of America, 62,* S99. doi : 10.1121/1.2016488

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P.K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *NeuroReport, 17*(10), 957-962. doi: 10.1097/01.wnr.0000223387.51704.89.

Ito, T., Tiede, M., & Ostry, D.J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Science USA, 106*(4): 1245–1248. doi: 10.1073/pnas.0810063106

Jusczyk, P.W., Copan, H., & Thompson, E. (1978). Perception by 2-month-old infants of glide contrasts in multisyllabic utterances. *Perception & Psychophysics*, *24*(6): 515-520.

Jusczyk, P.W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology, 23*(5): 648-654. doi: 10.1037/0012-1649.23.5.648

Jusczyk, P.W., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances in by 2-month-old infants. *Perception & Psychophysics, 23*(2): 105-109.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, 72, 379-389. doi: 10.1121/1.388081.

Kern, S., Davis, B. & Zink, I. (2009). From babbling to first words in four languages: Common trends, cross language and individual differences. *Becoming eloquent: Advances in the Emergence of language, human cognition and modern culture*, Hombert, J.M. & D'Errico, F. (eds), Amsterdam/Philadelphia, John Benjamins' Publishing Company, 205-232.

Kuhl, P.K., & Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577): 1138-1141. doi: 10.1126/science.7146899.

Kuhl, P.K., & Meltzoff, A.N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *Journal of the Acoustical Society of America, 100*, 2425–2438. doi: 10.1121/1.417951

Kuhl, P.K., & Meltzoff, A.N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3): 361-384. doi: 10.1016/S0163-6383(84)80050-8.

Kuhl, P.K., Ramirez, R.R., Bosseler, A., Lin, J.F.L., & Imada, T. (2014). Infants' brain responses to speech suggest Analysis by Synthesis. *Proceedings of the National Academy of Science USA, 111*(31): 11238- 11245. doi: 10.1073/pnas.1410963111.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255, 606–608.

Laurent, R., Barnaud, M.-L., Schwartz, J.-L., Bessière, P., & Diard, J. (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review*, *124*, 572-602. doi: 10.1037/rev0000069.

Levitt, A., Jusczyk, P. W., Murray, J., & Carden, G. (1988). Context effects in two-month-old infants' perception of labiodental/interdental fricative contrasts. Journal of Experimental Psychology: Human Perception and Performance, 14*(3),* 361-368. doi: 10.1037/0096-1523.14.3.361

Liberman, A.M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America, 29*: 117-123. doi: 10.1121/1.1908635.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1): 1-36. doi: 10.1016/0010-0277(85)90021-6.

Liberman, A.M., & Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences, 4*(5): 187–196. doi: 10.1016/S1364-6613(00)01471-6.

Locke, J., *Phonological acquisition and change*. New York: Academic Press, 1983

Lœvenbruck, H., Grandchamp, R., Rapin, L., Nalborczyk, L., Dohen, M., Perrier, P., Baciu, M. & Perrone-Bertolotti M. (2018). A cognitive neuroscience view of inner language: to predict and to hear, see, feel. In *Inner Speech: nature, functions, and pathology*, Peter Langland-Hassan & Agustín Vicente (eds.), Oxford University Press, 131-167.

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science, 219*(4590): 1347-1349. doi: 10.1126/science.6828865.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 131–149). Kluwer Academic.

Mahmoudzadeh, M., Dehaene-Lambertz, G., Fournier, M., Kongolo, G., Goudjil, S., Dubois, J., … & Wallois, F. (2013). Syllabic discrimination in premature human infants prior to complete formation of cortical layers. *Proceedings of the National Academy of Science USA, 110*(12): 4846-4851. doi: 10.1073/pnas.1212220110.

MacNeilage, P.F., & Davis, B.L. (2001). Motor mechanisms in speech ontogeny: phylogenetic, neurobiological and linguistic implications. *Current Opinion in Neurobiology, 11*, 696–700.

Majorano, M., Vihman, M.M., & DePaolis, R.A. (2014). The relationship between infants' production experience and their perception of speech. *Language Learning and Development, 10*(2): 179-204. doi: 10.1080/15475441.2013.829740.

Meister, I.G., Wilson, S.M., Deblieck, C., Wu, A.D., & Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, *17(19)*: 1692-1696. doi: 10.1016/j.cub.2007.08.064.

Mersad , K., & Dehaene-Lambertz, G. (2015). Electrophysiological evidence of phonetic normalization across coarticulation in infants. *Developmental Science*, *19(5):* 710-722. doi: 10.1111/desc.12325

Möttönen, R., Dutton, R., & Watkins, K.E. (2013). Auditory-motor processing of speech sounds. *Cerebral Cortex, 23*, 1190–1197. http://dx.doi.org/10.1093/cercor/bhs110

Möttönen, R., Järveläinen, J. Sams, M., & Hari, R. (2004). Viewing speech modulates activity in the left SI mouth cortex. *NeuroImage, 24*(3): 731-737. doi: 10.1016/j.neuroimage.2004.10.011.

Ojanen, V., Möttönen, R., Pekkola, J., Jääskeläinen, I.P., Joensuu, R., Autti, T., & Sams, M. (2005). Processing of audiovisual speech in Broca's area. *NeuroImage, 25*(2): 333-338. doi: 10.1016/j.neuroimage.2004.12.001.

Patri, J.F., Perrier, P., Schwartz, J.L., & Diard, J. (2018). What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework. *PLoS Comput Biol* 14(1): e1005942.

Patterson, M.L., & Werker, J.F. (1999). Matching phonetic information in lips and voice is robust in 4.5- month-old infants. *Infant Behavior and Development, 22*(2): 237-247. doi: 10.1016/S0163-6383(99)00003-X.

Patterson, M.L., & Werker, J.F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*(1): 93-115. doi: 10.1006/jecp.2001.2644.

Perani, D., Saccuman, M.C., Scifo, P., Anwander, A., Spada, D., Baldoli, C.,…, Friederici, A. (2011). Neural language network at birth. *Proceedings of the National Academy of Science USA, 108*(38): 16056-16061. doi: 10.1073/pnas.1102991108.

Pons, F., Lewkowicz, D.J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Science USA, 106*(26): 10598- 10602. doi: 10.1073/pnas.0904134106.

Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Science USA, 103*(20): 7865-7870. doi: 10.1073/pnas.0509989103.

Sato, M., Grabski, K., Glenberg, A.M., Brisebois, A., Basirat, A., Ménard, L., & Cataneo, L. (2011). Articulatory bias in speech categorization: evidence from use-induced motor plasticity. *Cortex, 47*(8): 1001-1003. doi: 10.1016/j.cortex.2011.03.009.

Schwartz, J.L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action Control Theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5): 336-354. doi: 10.1016/j.jneuroling.2009.12.004.

Schwartz, J.L., Boë, L.J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics, 25*, 255-286.

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language, 164*, 77-105. doi: 10.1016/j.bandl.2016.10.004.

Skipper, J.I., Van Wassenhove, V., Nusbaum, H.C., & Small, S.L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*(10): 2387-2399. doi: 10.1093/cercor/bhl147.

van Son, N., Huiskamp, T.M.I., Bosman, A.J., & Smoorenburg, G.F (1994). Viseme classifications of Dutch consonants and vowels. *Journal of the Acoustical Society of America, 96*, 1341-1355. doi: 10.1121/1.411324

Stevens, K.N. (1980). Acoustic correlates of some phonetic categories. *Journal of the Acoustical Society of America, 68*, 836–842. http://dx.doi.org/10.1121/1.384823

Stevens, K.N., & Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*, 1358–1368. http://dx.doi.org/10.1121/1.382102

Stevens, K., & Halle, M. (1967) Remarks on analysis by synthesis and distinctive features. In: Models for the perception of Speech and Visual Forms, Cambridge MA: MIT Press, 88-102.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-reading* (pp. 3–51). London: Lawrence Erlbaum Associates.

Sussman, H., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: the orderly output constraint. *The Behavioral and Brain Science*, *21*(2): 241–259. doi: 10.1017/S0140525X98001174.

Sussman, H. M., Hoemeke, K., & Ahmed, F. (1993) A cross-linguistic investigation of locus equations as a relationally invariant descriptor for place of articulation. *Journal of the Acoustical Society of America, 94*, 1256–68. http://dx.doi.org/10.1121/1.408178

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991) An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America, 90*, 1309–25. http://dx.doi.org/10.1121/1.401923

Vihman, M.M. (1993). Variable paths to early word production. *Journal of Phonetics, 21*(1): 61-82.

Vilain, A., Dole, M., Loevenbruck, H., Pascalis, O., & Schwartz, J.-L. (2019, March 19). PhonCat. https://osf.io/fzb59/

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130-145.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 7*(1), 49-63. doi : 10.1016/S0163-6383(84)80022-3

Wilson, S.M., Saygin, A.P., Sereno, M.I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature. Neuroscience, 7*(7): 701-702. doi: 10.1038/nn1263.

Yeung, H.H., & Werker, J.F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24*(5): 603-612. doi: 10.1177/0956797612458802.