Developmental Review 73 (2024) 101137

Contents lists available at ScienceDirect

Developmental Review

journal homepage: www.elsevier.com/locate/dr

Learning to imitate facial expressions through sound

Narain K. Viswanathan [a,*], Carina C.J.M. de Klerk [b], Samuel V. Wass [a], Louise Goupil [c]

[a] *School of Psychology, University of East London, United Kingdom*
[b] *Department of Psychology, University of Essex, United Kingdom*
[c] *Université Grenoble Alpes, CNRS, Laboratoire de Psychologie et Neurocognition, Grenoble, France*

ARTICLE INFO

Keywords:
Imitation
Facial imitation
Learning
Correspondence problem
Associative learning
Sensorimotor coupling
Infant auditory perception

ABSTRACT

The question of how young infants learn to imitate others' facial expressions has been central in developmental psychology for decades. Facial imitation has been argued to constitute a particularly challenging learning task for infants because facial expressions are perceptually opaque: infants cannot see changes in their own facial configuration when they execute a motor program, so how do they learn to match these gestures with those of their interacting partners? Here we argue that this apparent paradox mainly appears if one focuses only on the visual modality, as most existing work in this field has done so far. When considering other modalities, in particular the auditory modality, many facial expressions are not actually perceptually opaque. In fact, every orolabial expression that is accompanied by vocalisations has specific acoustic consequences, which means that it is relatively transparent in the auditory modality. Here, we describe how this relative perceptual transparency can allow infants to accrue experience relevant for orolabial, facial imitation every time they vocalise. We then detail two specific mechanisms that could support facial imitation learning through the auditory modality. First, we review evidence showing that experiencing correlated proprioceptive and auditory feedback when they vocalise – even when they are alone – enables infants to build audio-motor maps that could later support facial imitation of orolabial actions. Second, we show how these maps could also be used by infants to support imitation even for silent, orolabial facial expressions at a later stage. By considering non-visual perceptual domains, this paper expands our understanding of the ontogeny of facial imitation and offers new directions for future investigations.

Introduction

Facial imitation & the correspondence problem

Imitation can be defined as the execution of a behaviour that is topographically similar to a behaviour that the agent has just observed in another agent. Action A is topographically similar to action B when the configuration and movement of the effectors involved in action A are equivalent to those involved in producing action B, for example, when an agent claps with both hands after observing another agent do the same. Imitation is ubiquitous in interpersonal human communication (Chartrand & Bargh, 1999; Chartrand,et al., 2005), and is thought to play a crucial role in social cognition (Arnold & Winkielman, 2020), social affiliation

[*] Corresponding author at: University of East London, School of Psychology, Water Lane, London E15 4LZ, United Kingdom.
E-mail address: u2124122@uel.ac.uk (N.K. Viswanathan).

https://doi.org/10.1016/j.dr.2024.101137
Received 23 October 2023; Received in revised form 3 June 2024;

(Chartrand et al., 1999), and social learning (Heyes, 2021). Given the importance of this phenomenon, it is not a surprise that the question of how young infants learn to imitate others has been a central question in developmental psychology for decades (Slaughter, 2021; Jones, 2017; Meltzoff & Moore, 1997), with a surge of interest in recent years. Of particular interest is how infants learn to spontaneously copy others' facial expressions (Davis et al., 2021). This has been argued to constitute a particularly challenging learning task, because facial expressions are perceptually opaque (Brass & Heyes, 2005). Broadly, the argument goes like this: infants cannot directly see their own facial expressions, so when we consider the mechanism by which an infant might imitate someone else's facial expression, a "correspondence problem" arises: how could infants represent which pattern of muscle activation will make their facial expression match an observed facial expression if they do not have a representation of what their own facial expression looks like?

The literature on facial imitation development has suggested that this correspondence problem is solved in one of two ways. The first view, proposed by Meltzoff and Moore (1997), is that there is an inborn capacity of infants to map the observed acts of the other person onto one's own bodily acts perceived through proprioception, and that this mapping is supported by prior learning from self-generated movement patterns, including in utero movements. The second view, proposed by Brass et al. (2005) and Hayes (2016), is that the correspondence problem is solved by the formation of couplings between visual and motor representations of facial expressions through sensorimotor learning during dyadic interactions in which caregivers imitate their infants' facial expressions. In the first section of this paper, we provide an overview of these two theories, which are known as Active Intermodal Mapping (AIM; Meltzoff & Moore, 1997) and Associative Sequence Learning (ASL; Heyes & Ray, 2000). Note that when discussing the mechanisms involved in ASL, we will reference them as 'associative learning' in line with recent work by the authors who proposed the account (Heyes, 2021; Heyes, 2023).

We then go on to argue, in the second section, that these models are limited in scope because they mainly rely on empirical evidence that has focused on the visual modality (see Jones, 2007, and Isomura & Nakano, 2016, for two exceptions). While the current theories do not necessarily dismiss the involvement of other modalities (e.g., audition), theorising has largely focused on the visual domain when attempting to solve the correspondence problem for facial imitation. However, many facial expressions, in particular orolabial expressions (i.e., those involving the mouth and lips) have acoustic consequences that can be perceived by both the speaker and the listener. In other words, orolabial facial expressions can actually be perceptually transparent in the auditory modality when they are produced alongside vocalisations (Arias et al., 2018a; Arias et al., 2021a; de Gelder et al., 1999; Tartter, 1994; Quené et al., 2012). As such, these audible facial expressions are susceptible to associative learning, in the sense that when infants vocalise, they can accrue relevant contingent audio-motor experience allowing them to build sensorimotor schemes that can later support facial imitation.

In the third section, we then detail how a purely auditory route could support facial imitation learning for orolabial expressions (such as lip stretching or pouting) that are accompanied by vocalisations (Mechanism 1). We will first introduce the proposed mechanisms before describing how infants could build the audio-motor maps (or couplings) that can support the imitation of other's facial expressions purely through sounds. We review empirical evidence that shows how the relatively mature acoustic acuity observed in neonates, together with their sensitivity to the acoustic consequences of orolabial expressions, and the fact that infants extensively vocalise (even when alone) during the first few years of life, routinely provides them with exactly the type of contingent proprioceptive and auditory feedback they need to build audio-motor maps (or couplings) through associative learning.

Next, in the fourth section, we explain how audio-motor couplings could then be associated with audio-visual couplings when infants encounter opportunities to associate visual representations of specific facial expressions with their acoustic consequences, to support more diverse types of facial imitation, including the imitation of silent facial expressions. Finally, in the last section, we discuss the theoretical implications of these auditory and multimodal routes for our understanding of how infants learn to imitate facial expressions, and we outline directions for future research.

Theoretical accounts of the ontogeny of spontaneous facial imitation

Active intermodal mapping (AIM)

In the 70s and 80s Meltzoff and Moore reported evidence of facial imitation in neonates in the first few hours (Meltzoff & Moore, 1983) and weeks after birth (Meltzoff & Moore, 1977) and proposed the Active Intermodal Mapping (AIM) theory to explain this phenomenon (Meltzoff & Moore, 1997). AIM suggests that imitation is a matching-to-target process supported by two basic perceptual-motor processes that are present at birth: a proprioceptive feedback loop and a supra-modal sensory space. Comparison between action execution and observation is possible because both are coded within a supramodal representational system. The proprioceptive feedback loop enables the infant to compare the motor action to the observed target action, allowing them to correct the motor act to match the target (Meltzoff et al., 2017). When infants are engaged in a face-to-face interaction, there is a constant recurrence of proprioceptive feedback loops during which the infant actively compares the visual topography of the observed action with the topography of their own action in the supra-modal space. This allows the infant to change the topography of their own action to match the one they observed (Meltzoff et al., 1997; Meltzoff et al., 2017).

Yet, the existence of neonatal imitation is not settled science. Studies that reported evidence of neonatal imitation have been critiqued in recent years, owing to issues concerning replicability and statistical rigour (Davis et al., 2021; Ray & Heyes, 2011; Slaughter, 2021). Some papers report spontaneous facial imitation in neonates with moderate to large effect sizes (Meltzoff et al., 1977, 1983; Legerstee, M. 1991) while others (Anisfeld et al., 2001; Barbosa, 2017), including the study with the largest sample size and statistical power (Oostenbroek et al., 2016), found no evidence for neonatal facial imitation. The conclusions and analyses of the study with the highest power to date (Oostenbroek study, 2016) have been questioned in a follow-up paper (Meltzoff et al., 2018). They state that elements of the study design were ill-advised and attenuated infant imitation. Additionally, they published a reanalysis of the raw data that shows evidence of infant imitation for one behaviour (tongue protrusion; Meltzoff et al., 2018). A comprehensive discussion of this debate is beyond the scope of this paper (see Davis et al., 2021, for further discussion), but what is important for our aim here is that there is a lack of consensus regarding the presence of facial imitation in neonates. Furthermore, as we will see below, some evidence suggests that facial imitation is boosted by specific experiences, in particular interactions with caregivers (Rayson, et al., 2017; de Klerk, et al., 2019). Inborn predispositions are most plausible for behaviours that are present in neonates and relatively immune to experience. However, if facial imitation is not present in newborns and has a later onset time, and if it is related to infants' opportunities to experience specific contingencies in their environment, learning-based explanations – such as those proposed by the ASL account discussed in the next section – become more plausible.

Associative sequence learning

Associative Sequence Learning (ASL) has been proposed as an alternative to the AIM model, and specifies learning-based mechanisms that could support infants' learning of spontaneous facial imitation during the first few months of life (Heyes et al., 2001; Heyes, 2016). The only innate faculty that is required in these models is a simple domain-general associative learning mechanism (Heyes, 2001). The ASL model suggests that, when an agent consistently sees two temporally contingent and contiguous stimuli, a coupling between the neural representations of these two stimuli is formed as a result of Rescorla-Wagner learning (Rescorla, 1972; Cooper, et al., 2013). Rescorla-Wagner learning is similar to the Hebbian account of neural coupling ('neurons that fire together wire together') with the exception that non-contingent experience can reduce the coupling. This sensitivity to contingency has been validated by research that has shown changes in imitation behaviour as a consequence of training with varying contingencies and paired stimuli (Cooper et al., 2013). Caregivers and adults often imitate infants' facial expressions when engaged in face-to-face interaction (Moran et al., 1987; Rayson et al., 2017; de Klerk et al., 2019). According to ASL models, the acquisition of facial imitation depends on caregivers providing infants with this contingent, imitative feedback. This feedback is proposed to allow infants to form couplings between the visual, motor, and proprioceptive representations of their facial expressions (Ray et al., 2011). The ASL account suggests that, once coupled, the visual representation of a facial expression previously encountered alongside a specific motor-program will activate the associated motor response (Heyes, 2016, 2021; Slaughter, 2021).

One important consequence of this visuocentric route to facial imitation development is that infants would only be able to acquire correlated visual-motor experience for facial expressions during face-to-face exchanges where they can see another agents' facial expressions, or when placed in front of a mirror. This suggests that learning how to spontaneously imitate facial expressions would *require* direct face-to-face dyadic interactions (or access to a mirror), and that it would be highly dependent on the consistency of the partner's tendency to copy the infant during those interactions. Recent work has provided evidence for this: mothers' tendency to imitate their infants' facial expression has been shown to predict both the infant's tendency to imitate the facial expressions of a video model (as measured by increased EMG activity in the corresponding muscle regions; de Klerk et al., 2019) and increased motor system activation during the observation emotional facial expressions (Rayson et al., 2017). This supports the idea that parental imitation can allow infants to build visuomotor couplings that support imitation.

Yet, ASL models of facial imitation have been critiqued for this dependency on dyadic exchanges. First, observers of child-parent interactions have reported that when caregivers imitate the infant, they usually display exaggerated versions of the infant's expression (Brand, Baldwin, & Ashburn, 2002; Meltzoff et al., 2017). If imitation is a consequence of ASL, then facial imitation should primarily be observed when infants observe exaggerated facial expressions, and they would only display a mismatched, subdued version of the observed expression in response. However, it is important to note that there is no evidence to suggest that facial mimicry is well matched even in adults – i.e. many studies do not observe overt mimicry in response to facial expressions but measure subthreshold EMG activity (e.g., McIntosh et al., 2006). Second, according to the principles of ASL, instances where the infant's execution of a particular expression is not followed by the observation of the same expression enacted by the adult will reduce the strength of any coupling. If the correlated co-occurrences of both stimuli increase the likelihood of coupling, then instances of absent co-occurrence would reduce the likelihood that the two stimuli will be effectively coupled. Research shows that during face-to-face interactions, specific infant facial expressions are mirrored 30% to 55% of the time (Rayson et al., 2017; de Klerk et al., 2019), suggesting that many facial expressions may receive more non-contingent than contingent feedback, even when executed in a dyadic context (Meltzoff et al., 2017). When we consider this in combination with the fact that infant is likely to express the same facial expression when alone, it becomes possible that the Rescorla-Wagner couplings associated with those facial expressions have opportunities to weaken in strength because of the frequent absence of contingent feedback.

There is an additional factor that suggests that some of the contingent feedback that infants perceive in their first few months of life may not be of the highest quality: poor visual acuity. Indeed, infants have poor visual acuity until the age of 4 months: infants' visual acuity at birth is 1/40th of normal vision (Maurer & Lewis, 2001) and is less than 1/20th of normal vision until they reach three months (Chandna, 1991). This suggests that the actual visual percept of the partner's face itself changes over the first months of life because of improving visual acuity. Until visual acuity has matured and stabilised, visual-motor couplings formed at different levels of visual acuity might not be activated to the same degree when the visual representation itself changes owing to improving acuity. As we will discuss below, infants' auditory acuity is comparatively more developed, which suggests that audio-centric sensorimotor couplings do not have the same limitation.

In sum, visuocentric versions of ASL theory primarily rely on dyadic exchanges and the perception of contingency in the visual modality to explain how infants learn to imitate facial expressions. As we have seen, empirical evidence suggests that parental feedback experienced in the visual modality is indeed an important factor predicting infants' imitation (de Klerk et al., 2019; Rayson et al., 2017). Yet, the frequencies with which parents actually imitate their infants, and infants' poor visual acuity, may limit their ability to learn purely through the visual modality, and it therefore seems important to explore the contribution of other modalities, in particular perceptually transparent ones like audition.

In the following sections, we describe how auditory routes to facial imitation can be established without the need for interpersonal contingency. Routes that are not systematically dependent on dyadic exchanges, including the purely auditory one we describe here, could play a complementary role to visual routes grounded in dyadic exchanges. In particular, they would allow infants to form perceptuomotor couplings for orolabial facial gestures such as mouth opening and smiling that are strenghtened when infants accrue sensory experience while exploring their environment outside of contingent social interactions.

Could the auditory modality provide an additional route to the acquisition of facial Imitation?

Facial expressions that involve the mouth and lips (orolabial expressions) often have acoustic consequences that are perceptible to the speaker and listener when they are produced alongside speech or vocalisations. Essentially, the mouth acts as a kind of "filter", and depending on the shape and size of the mouth cavity, relative formant frequencies and amplitudes of vocal sound waves are modified (Titze, 1994). This relationship between mouth shape/size and spectro-temporal features is discussed in greater detail in 'The relative transparency of orolabial facial expressions' below. The existence of this link would suggest that certain orolabial facial expressions are not fully opaque in the auditory modality; i.e. change in mouth cavity shape/size causes changes in formant frequencies that are perceptible to the speaker themselves. Furthermore, auditory acuity is relatively high during the first few months of life, and infants are sensitive to these perceptual features relatively early in development, as the evidence below will illustrate. We therefore suggest that the auditory domain may offer an important, complementary route that can support infants' facial imitation development of certain orolabial facial gestures (e.g. pouting, smiling). From an ASL perspective, this opens up the possibility that infants might be able to accrue correlated perceptual-motor experience relevant for facial imitation learning for these gestures even while alone, i.e., without the need for dyadic exchange.

The perceptually transparent auditory route

Jones (2017) proposed that during early infancy, infants do not imitate visually perceived expressions; instead, they initially try to replicate and reproduce sounds that they hear. This proposal was supported by a longitudinal study (Jones, 2007) that tracked infant's imitation of actions modelled by their parents. The imitation of 'Aah' vocalisation was observed at an earlier age when compared to visual, silent behaviours. While this account did not seek to address or explain how infants can learn to imitate facial expressions *per se*, it provides support for the idea that audition supports imitation learning during infancy. Similarly, Kuhl & Meltzoff (1996) introduced

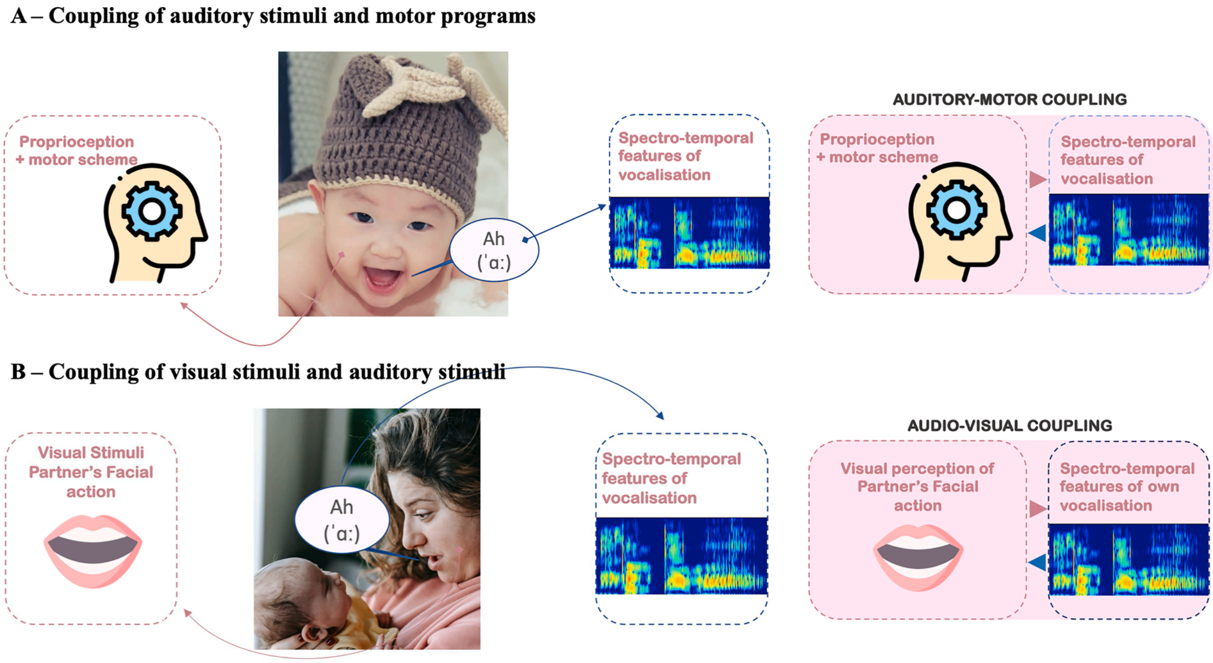
A – Coupling of auditory stimuli and motor programs
Proprioception + motor scheme
Ah ('ɑ:)
Spectro-temporal features of vocalisation
AUDITORY-MOTOR COUPLING
Proprioception + motor scheme
Spectro-temporal features of vocalisation
B – Coupling of visual stimuli and auditory stimuli
Visual Stimuli Partner's Facial action
Ah ('ɑ:)
Spectro-temporal features of vocalisation
AUDIO-VISUAL COUPLING
Visual perception of Partner's Facial action
Spectro-temporal features of own vocalisation

Fig. 1. Depiction of the correlated sensory stimuli during different context. A) When the infant babbles, there is a correlation between the motor program that has been executed and the spectro-temporal features of the sound that is produced. B) A depiction of the correlated sensory stimuli when the caregiver vocalises while executing an orolabial expression. The infant receives correlated exposure to the acoustic consequence of the facial expression and the visual perception of the expression.

'auditory-motor maps' that are formed by associations between "self-produced auditory events" and "the motor movements that caused them" (p. 2436; NB: this is similar to the audio-motor couplings that we will describe below, see Fig. 1A). This mechanism was introduced as a potential building block for speech acquisition, aiming to explain how infants learn to imitate vocal sounds. For a detailed discussion of auditory-motor associations involved in vocal development theories, see Messum et al. (2015).

Building on the more general framework of ASL, here we suggest that early facial imitation can be acquired based on such sensorimotor couplings. We describe hypothetical mechanisms through which audio-motor maps that were initially built in the auditory domain by associative learning can later be used to copy most orolabial expressions, including silent ones (such as smiles). Positing the formation of couplings through associative learning allows us to sketch out mechanisms that naturally follow from the implications of ASL without the need for any additional conjecture to explain how infants might learn to imitate facial expressions through sounds.

In the following, we will first describe how such a purely auditory learning route to spontaneous facial imitation might work. We then discuss how infants might learn the required audio-motor couplings, and describe evidence that supports the existence of such audio-motor couplings during early infancy.

Audio-motor coupling – Mechanism 1 – A purely auditory route for spontaneous facial imitation

Mechanism 1 describes how facial imitation of orolabial gestures could arise when an infant is engaged in an interaction with an adult who produces a sound that has a specific spectro-temporal signature. This mechanism is enabled by audio-motor maps that can be built by the infants through vocalising, even when alone (Fig. 1A). When the infant babbles, there is a correlation between the motor program that has been executed and the changes in spectro-temporal features of the sound that is produced. When adult vocalisations contain spectro-temporal changes that match these acoustic characteristics, the auditory perception (2A, Fig. 2) of the vocalisation (1B, Fig. 2) that accompanies the caregiver's facial expression could trigger the activation of a pre-existing coupled audio-motor representation in the infant brain (2B, Fig. 2). If this activation reaches a certain threshold it may result in the execution of a facial expression that is similar in topography to the one produced by the adult (3, Fig. 2). This constitutes a purely auditory route where

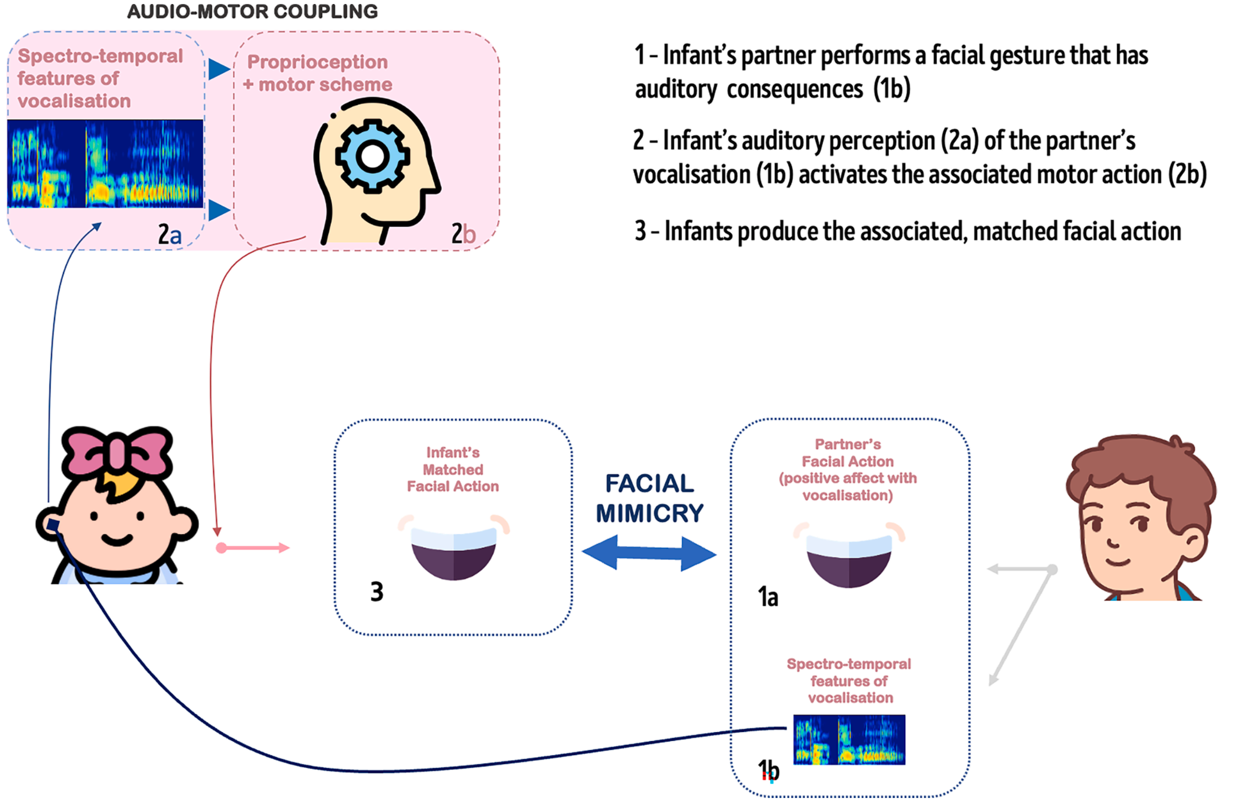
AUDIO-MOTOR COUPLING
Spectro-temporal features of vocalisation
Proprioception + motor scheme
2a
2b
1 - Infant's partner performs a facial gesture that has auditory consequences (1b)
2 - Infant's auditory perception (2a) of the partner's vocalisation (1b) activates the associated motor action (2b)
3 - Infants produce the associated, matched facial action
Infant's Matched Facial Action
3
FACIAL MIMICRY
Partner's Facial Action (positive affect with vocalisation)
1a
Spectro-temporal features of vocalisation
1b

Fig. 2. Mechanism 1 – Schematic representation of how facial imitation can be supported by the proposed audio-motor coupling. Facial imitation can arise through this route when an infant is engaged in an interaction with an adult who produces a sound that has a specific spectro-temporal signature. If the infant has already built an audio-motor map through vocalising alone that contains an exemplar that matches these acoustic characteristics, the auditory perception (2a) of the vocalisation (1b) that accompanies the caregiver's facial expression triggers the activation of the pre-existing coupled motor representation (2b). If this activation reaches a certain threshold it may result in the execution of a facial expression that is similar in topography to the one produced by the adult (3).

audio-motor maps could activate facial imitation solely based on the presentation of auditory information.

Audio-motor couplings, and other sensorimotor couplings formed by associative learning, can be categorised as forward models, i.e., models that encode the probable sensory consequences of a programmed gesture (Prinz, 2005). For infants to be able to form such audio-motor models while vocalising (even when alone), the following conditions would need to be fulfilled:

1) Orolabial expressions must have reliable and consistent acoustic consequences.
2) Infants must be able to hear and discriminate the relevant spectro-temporal features of their own vocalisations to associate them with specific motor programs.

Below we review evidence supporting both of these claims.

The relative transparency of orolabial facial expressions

During vocal production, the mouth cavity acts as a resonator for sounds that are generated by the vocal folds. When a person changes the shape and size of their mouth, this modulates the spectro-temporal structure of the sound produced (Titze, 1994). In particular, when one smiles, the mouth both expands and widens, resulting in an increase in the frequencies of the first two formants (peaks in spectrogram of the vocalisation) and an increase in amplitude of the third formant regardless of the phoneme that is being produced (Arias et al., 2018b; El Haddad et al., 2015; Podesva et al., 2015; El Haddad et al., 2017; Drahota et al., 2008; Tartter, 1980). Conversely, pouting causes a lowering of the first formants (Tartter & Braun, 1994), and vowels vocalised while holding a disgusted face lower the first two formants (Chong, Kim & Davis, 2018). Of course, other articulators, including the jaw and the tongue, also change the spectro-temporal features of vocalisations. For example, when the extent to which the jaw is open is kept constant and the tongue alone is moved from a forward to a backward position, the sound of vocalisations shift from an ‘i’-like sound to an ‘u’-like sound (Morrish et al., 1985). In the remaining however, we only focus on *visible articulators*, because we are concerned with the imitation of visible facial expressions.

The evidence reviewed above concerns adult speakers, but what about infants? Are these dependencies between the configuration of the main articulators and spectro-temporal features already present early in life? What is the evidence that infants are able to perceive and discriminate these acoustic features when they vocalise? Is there evidence that infants regularly vocalise with varying facial expressions? And what is the evidence that these experiences are sufficient to enable infants to build forward models (i.e. models that encode the probable sensory consequences of an gesture) linking specific articulatory motor programs with resulting auditory percepts?

Changes in the spectro-temporal features of sound are perceptible to infants

A first key challenge for our proposal is whether infants are able to hear and discriminate the spectro-temporal features of the vocalisations they themselves produce. In recent years, evidence has been accumulating for the idea that infants are born with, or quickly acquire, sophisticated auditory abilities. For example, research has shown that neonates and even third-trimester fetuses show a differential response to their mother’s compared to their father’s (Lee & Kisilevsky, 2014) or a female stranger’s (DeCasper & Fifer, 1980; DeCasper et al., 1994; Kisilevsky et al., 2003, 2009; Moon et al., 1993) voice. This capacity to discriminate specific voices, especially when considering the high amplitude endogenous noise that the fetus is exposed to (Parga et al., 2018), suggests that infants are proficient at processing spectro-temporal information even before birth. During early infancy, infants’ “perceptual sensitivities” (Eimas et al., 1971; Werker & Tees, 1999) are wider than that of adults, before narrowing down towards the end of the first year, a process known as perceptual narrowing. For instance, one-month old infants show the capacity to make speech-sound distinctions that are not observed in older infants (Eimas et al., 1971). In parallel with this finding, infants’ capacity to discriminate between phonemes of non-native languages reduces between 6 months to 12 months showing that there is a narrowing in the type and number of phonemes that the infants are sensitive to (Kuhl et al., 2006; see Werker et al., 2012 for review of this literature). More recently, EEG studies have confirmed that neonates are already sensitive to phonetic contrasts (Cheng et al., 2012; Dehaene-Lambertz & Pena, 2001; Mastropieri & Turkewitz, 1999). This suggests that during early infancy, infants’ auditory perception may be even more sensitive to subtle spectrotemporal features than that of adults.

Infants experience a large amount of correlated audio-motor experience

A second key challenge for our proposal is whether infants’ articulatory skills are sufficient for them to produce vocalisations that are sufficiently diverse so as to support the acquisition of forward models linking specific motor programs with their acoustic consequences.

Infants have a strong tendency to explore their vocal abilities. All-day recordings in the home and lab studies show that infants produce about four vocalisations per minute from the first month of life (Oller et al., 2013, 2019; Iyer et al., 2016), which provides them with numerous instances in which motor activation and proprioceptive feedback occur in synchrony with specific auditory stimuli. The majority of infant vocalisations are protophones which are speech-like sounds involving vocants (vowel-like), squeals or growls (Oller et al., 2013, suppl.). In addition to protophones, all-day home recordings have shown that cries occur at the rate of approximately 0.5 incidences per minute from the neonatal period and laughter occurred at the same rate as crying from 3 months onwards (Oller et al., 2021). Furthermore, studies examining infants’ vocalisations rates when their caregiver is inside or outside of their visual field suggest that infants vocalise both when the caregiver is attentive to them and when they are disengaged (e.g., interacting with another agent; Long et al., 2020).

These vocal explorations provide infants with relevant experiences to build forward models, because the above-mentioned dependencies between the configuration of the main articulators and the spectro-temporal features of vocalisations are already present early in life (Serkhane et al., 2007; Choi et al., 2021). Over the first few years of life, the acoustic space of vocalisations narrows down to progressively reflect children's linguistic environment (Vorperian & Kent, 2007; Serkhane et al., 2007), but even the earliest vocal explorations are not fully random. They are already constrained by infants' linguistic environment (Mampe, Friederici, Christophe, & Wermke, 2009; Kisilevsky et al., 2009; Levitt & Wang, 1991), but also by anatomical (Serkhane et al., 2007) and physiological (Wass et al., 2022). factors. In particular, vocal explorations are initially constrained by articulatory development, which evolves over the first few months and years of life, with an increasing involvement of certain articulators like the jaw (Serkhane et al., 2007), while movements of the lips remain very variable during the first two years of life (Green, Moore & Reilly, 2002). To summarise, what we know from these studies is that, over the course of the first two years of life, infants execute motor programs that produce vocalisations that are acoustically varied. Given that mouth cavity size (and shape) affects the spectrotemporal features of the sounds they produce, when they vocalise infants can perceive large and diverse amounts of correlated auditory-proprioceptive feedbacks.

We also know that, at a broad level, there is a high level of consistency in the co-occurrence of specific vocalisation types and accompanying facial expressions (and corresponding motor activation and proprioceptive feedback). For instance, Oller et al. (2013) tracked the valence of infants' facial expression during cries and laughter. They identified that cries typically co-occur with highly negative facial affect expressions and laughter always co-occurred with highly positive facial affect expressions (i.e., smiles). This suggests that there are types of vocalisations that almost always co-occur with specific orolabial arrangements, e.g. infants repeatedly experience the characteristic proprioceptive feedback resulting from stretching or pouting their lips alongside vocalisations that have higher (or lower) formant frequencies. A higher quantity of experience for these correlated stimuli would suggest that sensorimotor couplings for these stimuli might be established earlier to other correlated stimuli.

Recent articulatory impairment studies suggest that these experiences are sufficient to enable even 3-month-old infants to link specific orolabial motor programs with auditory percepts (Bruderer, et al., 2015; Choi et al., 2021). For example, a study found that inhibiting tongue-tip movements in 3-month-old infants disrupted infants' ability to discriminate between two phonetic stimuli that involve these articulators, as evidenced through ERP responses (Choi et al., 2021). The phonetic stimuli used in this comparison required orolabial expressions that were identical, with the exception of the timing and placement of tongue-tip movements. If a coupling between the tongue-tip gesture and the auditory consequence had not been established, a differential ERP response would not have been observed. Thus, by 3 months of age, infants appear to have already formed perceptual-motor couplings (or forward models) that link together motor programs and auditory percepts. Although the existing evidence involved other articulators than the mouth, it is plausible that this holds true for visible articulators as well.

Direct evidence for the existence of a purely auditory route to facial imitation in adults

In the two preceding sections, we have shown that infants' vocal explorations and auditory abilities allow them to build forward models linking specific motor programs and their associated auditory percepts. But could these early audio-motor couplings support facial imitation of orolabial gestures via a purely auditory route during infancy? This possibility is supported by recent papers which show that adults' facial muscle responses can be based solely on changes in spectro-temporal features of vocalisations. Arias and colleagues (2021a) increased the first two formant frequencies of recorded human vocalisations while leaving all other aspects (e.g., intonation) unchanged. The formant structure was changed in a manner that replicated the effect of smiling (larger, wider mouth cavity; Arias, 2021b). When cortically blind participants listened to these stimuli, increased EMG activity was observed over their zygomaticus major – the muscle responsible for lip-pulling and smiling. This suggests that there is a coupling between the auditory stimuli that the participants heard and the motor programme that was activated, i.e., there is a direct audio-motor route linking the smiling sound and the gesture of smiling. This demonstrates that purely auditory routes to facial imitation do exist during adulthood, but whether these routes are already functional during infancy remains to be demonstrated.

Acquired equivalence and the audio-visual–audio-motor, throupled route.

In this section we now discuss the issue of how infants might be able to use audio-motor couplings to learn to imitate even *silent* facial expressions. We suggest that this can happen if infants are able to match the coupled audio-motor couplings discussed above (Fig. 1A) with additional audio-visual couplings that can be built during dyadic interactions (Fig. 1B). During instances where infants are engaged in a face-to-face interaction with their caregiver and are able to visually perceive changes in their caregiver's face (Fig. 1B), they will be exposed to instances where caregivers are producing multiple orolabial configurations. This provides the infant with the opportunity to associate the visual representation of specific orolabial configuration changes with the corresponding change in spectro-temporal features, allowing them to build audio-visual coupling. If the infant has exposure to similar changes in spectro-temporal features while they are babbling alone and while observing their caregiver, those spectro-temporal features will then be present in two different couplings: the audio-motor coupling and the audio-visual coupling. When two separate couplings both include the same or similar stimuli, the two separate couplings could combine, resulting in an *acquired equivalence* between the couplings and their stimuli (Fig. 3; Heyes, 2016). Once this is established, it may be possible for the newly *throupled* coupling to enable infants to perform facial imitation of these orolabial gestures based on visually perceived changes in face configuration alone (Fig. 3, below). The existence of this route would thus enable the infant to imitate the facial expression even when it is not accompanied by a vocalisation.

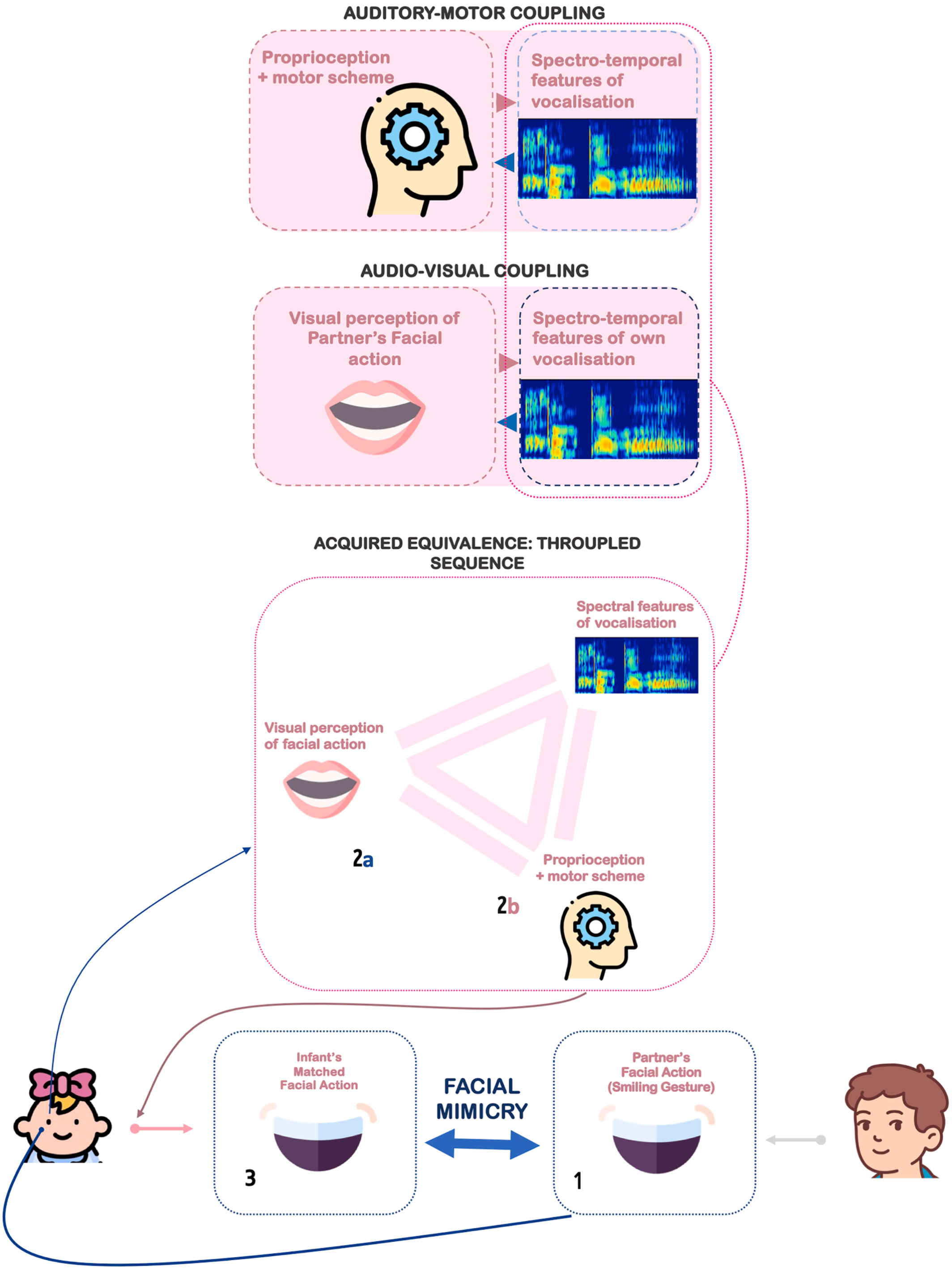
AUDITORY-MOTOR COUPLING
Proprioception + motor scheme
Spectro-temporal features of vocalisation
AUDIO-VISUAL COUPLING
Visual perception of Partner's Facial action
Spectro-temporal features of own vocalisation
ACQUIRED EQUIVALENCE: THROUPLED SEQUENCE
Spectral features of vocalisation
Visual perception of facial action
2a
Proprioception + motor scheme
2b
Infant's Matched Facial Action
3
FACIAL MIMICRY
Partner's Facial Action (Smiling Gesture)
1

Fig. 3. Mechanism 2 – Combining audio-visual and audio-motor couplings to learn to imitate silent facial expressions. Schematic of facial imitation that is facilitated by the throupled, acquired equivalence between the two couplings. An illustration of what happens when an infant is engaged in an interaction with an adult who produces a facial expression whose acoustic consequence is present in two separate couplings: a coupled audio-visual coupling & a coupled audio-motor coupling. The visual perception (2a) of the expression (1) that accompanies the facial expression automatically triggers the activation of a throupled motor representation (2b). If this activation reaches a certain threshold it may result in the execution of a facial expression that is similar in topography to the one produced by the adult (3).

Research has shown that by three months, infants are able to both identify different phonemes (perceptually different units) and form couplings with visual stimuli that are presented with them (Mersad, Kabdebon, & Dehaene-Lambertz, 2021) suggesting that infants can form couplings between auditory and visual stimuli cross-modally. Evidence that infants possess audio-visual couplings can be found in the McGurk paradigm literature. McGurk paradigms manipulate the auditory stimuli that are paired with each visual stimulus. Differential responses to varying audio-visual stimulus pairs (that differ in real-world co-occurrence rates) indicates that there is variance in the degree of pairing between the visual and auditory stimuli. A difference in response would suggest that one of the pairs has been experienced together more consistently and that there is a coupling between the pairs that have been experienced more consistently. This differential response has been observed in multiple studies (Bahrick & Lickliter, 2004; Bristow et al., 2008; Kuhl & Meltzoff, 1982,1984; Kushnerenko et al., 2008; Patterson & Werker, 2003; Rosenblum et al., 1997; Yeung & Werker, 2013), showing that by 5-months infants have developed a coupling between many sounds, and the visual perception of the orolabial gesture necessary to produce those sounds. When audio and visual stimuli are simultaneously presented, like they are in the McGurk studies, they are referred to as bimodal stimuli. Kuhl & Meltzoff (1984) postulated that the differential response to bimodal stimuli was proof that speech sounds are "stored" in a multimodal manner in infants. Adding further support, a differential response to bimodal stimuli has also been observed in affective development literature: infants are able to identify mismatches in the emotional category of audio and visual stimuli by the age of 4 months (Flom & Bahrick, 2007). For example, respond differently for happy speech accompanied by happy facial expression when compared to happy speech accompanied by an angry facial expression. This capacity to discriminate between emotional category of audio-visual stimuli evidences the presence of audio-visual couplings between the visual features and auditory features of affect (emotion). We have explored converging evidence from the fields of phoneme discrimination, McGurk effect and affective discrimination that support the existence of audio-visual couplings in infants from at least 4 months of age.

Discussion

Existing theories concerning facial imitation learning – AIM and ASL – do not explicitly exclude the involvement of the auditory realm. ASL is explicitly domain-general, while the AIM model defines a *supra* modal sensory space. However, discussions of these theories in the literature tend to focus on the visual modality and largely disregard the auditory one (Heyes, 2016; Meltzoff et al., 2017). This is likely due to the fact that the available empirical evidence primarily concerns the visual modality. Here, we suggest that associative learning in the auditory modality could be a powerful mechanism allowing infants to learn how to imitate orolabial facial expressions, and have reviewed relevant evidence supporting this claim. This overview also highlights the fact that more research focusing on the auditory realm is needed to examine the existence of purely auditory routes to the imitation of many orolabial gestures. We hope that this article will inspire researchers to move in this direction, and below, we review several perspectives opened up by our proposal, and suggest directions for further research.

Typically, facial expressions have been thought of as perceptually opaque, and as a result, it has been proposed that spontaneous facial imitation should be observed at a later age than the spontaneous imitation of perceptually transparent expressions such as hand movements and vocalisations (Heyes, 2021). All that is necessary for facial imitation is the capacity to reproduce a perceived expression using facial muscles. The auditory route to spontaneous facial imitation we describe above would suggest that infants can accrue perceptual-motor couplings required for facial imitation while vocalising alone, without the need for dyadic interactions, for facial expressions that are typically accompanied by vocalisations. The implication of the models we proposed is that voiced facial expressions are not perceptually opaque, so if the mechanisms we described in this paper exist, we would expect specific orolabial expressions to be imitated as soon as infants are capable of producing the corresponding facial gestures. This would also suggest that infants may be able to copy orolabial gestures such as smiling, mouth opening, and pouting at an earlier age than facial gestures, such as eyebrow raising or frowning, for which they can only develop perceptual-motor couplings via a visual route. Currently, no strong empirical evidence is available to directly test this assertion since no study has directly looked at infant spontaneous facial imitation to diverse audio stimuli (when visual information is unavailable) in a longitudinal design.

Our proposal that infants are able to accrue the relevant perceptual-motor experience while alone also counters some of the criticisms of visuocentric models of associative learning. It has been pointed out that all of the instances in which an infant performs a facial expression without observing the same facial expression (on their own or someone else's face) will weaken their perceptual-motor couplings linking specific visual percepts with the corresponding proprioceptive feedbacks and motor commands (e.g., Meltzoff et al., 2017). One way to argue against this is to consider that context is a core feature of ASL: in this view, a coupling built during a face-to-face dyadic context would not be eroded in other (non-dyadic) contexts (Heyes, 2016). Still, it is worth noting that this issue would not even arise if the couplings did not strictly rely on dyadic exchanges, as is the case for the purely auditory route we described here.

There is a possibility that associated affects also enhance the imitation observed. For example, it is possible that hearing an increase in formant frequencies that accompany a smile also evokes a feeling of happiness. This would mean that there is an additional factor – an affective state – that is associated with the coupling present. In other words, the association between sounds and gestures might be partially, or completely, mediated by affects.

An additional criticism that has been linked to the visual-motor associated couplings experienced in the dyadic context is that caregivers and adults do not attempt a perfect imitation of infant's facial expression. Conventionally, it is thought that adults display an exaggerated version of the infant's facial expression, possibly to hold the infant's attention for longer and/or to elicit positive affect (Meltzoff et al., 2017). Therefore, the correlated visual-motor coupling will not involve a motor program and visual stimulus that are similar to each other, i.e., the infant will produce a much more understated, not wholly similar version of the observed expression. Because our proposed routes do not rely on couplings that are established during dyadic exchanges and interpersonal contingency, they are not susceptible to this criticism. Both of the couplings that are discussed in this paper (Fig. 1, above) involve correlated stimuli that are perfectly matched to each other as they are both produced by the same actor as the same time.

A limitation of our account is that the auditory route we described will only enable imitation of orolabial actions that *are accompanied with vocalisations*. Some of the facial actions that are studied in the facial imitation literature (e.g. tongue protrusion [Meltzoff et al., 1977], eyebrow movement [de Klerk et al., 2019]) do not have auditory consequence, and as such they cannot be explained through the mechanism we have laid out. Another limitation of this account is that it remains to be fully validated empirically. Below we lay out several key venues for future research aiming to test this model.

First, if our proposed Mechanism 1 is valid, then just providing auditory information of which the spectro-temporal features are associated with an orolabial gesture (e.g. a recording of the infant’s own babble that occurred during a smile-like facial affect) should be sufficient to trigger the infant’s production of the same gesture. Manipulating the spectro-temporal features of presented audio-visual orolabial expression stimuli should change the probability that infants would imitate this facial expression if these spectro-temporal features are sufficient for activation of the coupled motor program. If a variance in facial imitation response is observed as a consequence of this manipulation, it would suggest that audio-motor couplings play a role in enabling orolabial imitation even when visual information is available. Relatedly, Mechanism 1 would also predict that the orolabial, facial imitation response of children with typical hearing should be superior to those of children with hearing impairments. Mechanism 1′s reliance on audio-motor maps built during babbling implies that infants’ babbling frequency overall (that is, regardless of caregivers’ contingency, Long et al., 2020) and audio-motor representations (Choi et al., 2021) should be associated with orolabial, facial imitation response and performance. Examining the direct predictions of our model will shed light on both the role of auditory realm in the development of facial imitation of orolabial gestures and the scope of ASL itself, as Mechanism 1 is a direct implementation of this theory in the auditory modality.

Mechanism 2 aims to explain the imitation of the orolabial expressions that have been conventionally focused on in this area of research: gestures produced without an accompanying vocalisation, i.e., where only visual information is available. Our mechanism predicts that infants should be able to imitate silent facial expressions based on visual stimulus only after they are able to reproduce orolabial, facial gestures in response to auditory stimuli that have specific spectro-temporal features. In other words, in a longitudinal design, Mechanism 2 should manifest at a later age to Mechanism 1 owing to the former’s need for a two couplings and their acquired equivalence. Evidence in support of Mechanism 2 would suggest that audio-motor couplings similar to visual-motor couplings can enable facial imitation of orolabial expressions in instances where only visual information is available and show the robust portfolio of faculties that ASL can provide.

Conclusion

In this paper, we have highlighted a missing piece in the facial imitation literature: how *audition* could support infants’ acquisition of the capacity to imitate orolabial, facial gestures, sketching out two different mechanisms by which this capacity might be learned during the first year of life, supporting first the imitation of voiced orolabial expressions, and potentially later on in development, the imitation of even silent orolabial gestures. The auditory realm has been largely ignored in this area, and our paper both highlights its relevance, and provides novel mechanisms that provide testable hypotheses for future research.

Funding information

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 853251 – ONACSA; https://doi.org/10.3030/853251).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The figures used in the paper have benefitted from the open licenses provided by Flaticon.com and Pexels.com. Figs. 1, 2 and 3 employ modified versions of images and icons made by Freepik and Victoruler from flaticon.com. Fig. 1 includes picture obtained from Pexels.com.

References

Anisfeld, M., Turkewitz, G., Rose, S. A., Rosenberg, F. R., Sheiber, F. J., Couturier-Fagan, D. A., … Sommer, I. (2001). No compelling evidence that newborns imitate oral gestures. *Infancy, 2*(1), 111–122.

Arias, P., Belin, P., & Aucouturier, J. J. (2018a). Auditory smiles trigger unconscious facial imitation. *Current Biology, 28*(14), R782–R783.

Arias, P., Soladie, C., Bouafif, O., Roebel, A., Seguier, R., & Aucouturier, J. J. (2018b). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing, 11*(3), 507–518.

Arias, P., Bellmann, C., & Aucouturier, J. J. (2021a). Facial mimicry in the congenitally blind. *Current Biology, 31*(19), R1112–R1114.

Arias, P., Rachman, L., Liuni, M., & Aucouturier, J. J. (2021b). Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review, 13*(1), 12–24. https://doi.org/10.1177/1754073920934544

Arnold, A. J., & Winkielman, P. (2020). The mimicry among us: Intra-and inter-personal mechanisms of spontaneous mimicry. *Journal of Nonverbal Behavior, 44*(1), 195–212.

Bahrick, L. E., & Lickliter, R. (2004). Infants' perception of rhythm and tempo in unimodal and multimodal stimulation: A developmental test of the intersensory redundancy hypothesis. *Cognitive, Affective, & Behavioral Neuroscience, 4*, 137–147.

Barbosa, P. G. (2017). Are You Like Me? Maybe, But I Will Not Imitate You! A Longitudinal Study on Newborns and Infants' Imitation and Conspecific Identification Skills (Doctoral dissertation, Doctoral dissertation. University of Alberta). Doi: https://doi.org/10.7939/R3C24R22F.

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': Modifications in mothers' infant-directed expression. *Developmental Science, 5*(1), 72–83.

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem? *Trends in Cognitive Sciences, 9*(10), 489–495.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J. F. (2008). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience, 21*(5), 905–921.

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences, 112*(44), 13531–13536.

Chandna, A. (1991). Natural history of the development of visual acuity in infants. *Eye, 5*(1), 20–26.

Chartrand, T. L., Maddux, W. W., & Lakin, J. L. (2005). Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. *The New Unconscious*, 334–361.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893.

Cheng, Y., Lee, S. Y., Chen, H. Y., Wang, P. Y., & Decety, J. (2012). Voice and emotion processing in the human neonatal brain. *Journal of Cognitive Neuroscience, 24*(6), 1411–1419.

Chong, C. S., Kim, J., & Davis, C. (2018). Disgust expressive speech: The acoustic consequences of the facial expression of emotion. *Speech Communication, 98*, 68–72.

Choi, D., Dehaene-Lambertz, G., Peña, M., & Werker, J. F. (2021). Neural indicators of articulator-specific sensorimotor influences on infant speech perception. *Proceedings of the National Academy of Sciences, 118*(20), Article e2025043118.

Cooper, R. P., Cook, R., Dickinson, A., & Heyes, C. M. (2013). Associative (not Hebbian) learning and the mirror neuron system. *Neuroscience Letters, 540*, 28–36.

Davis, J., Redshaw, J., Suddendorf, T., Nielsen, M., Kennedy-Costantini, S., Oostenbroek, J., & Slaughter, V. (2021). Does neonatal imitation exist? Insights from a meta-analysis of 336 effect sizes. *Perspectives on Psychological Science, 16*(6), 1373–1397.

Dehaene-Lambertz, G., & Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport, 12*(14), 3155–3158.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208*(4448), 1174–1176.

DeCasper, A. J., Lecanuet, J. P., Busnel, M. C., Granier-Deferre, C., & Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant Behavior and Development, 17*(2), 159–164.

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication, 50*(4), 278–287.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303–306.

El Haddad, K., Dupont, S., d'Alessandro, N., & Dutoit, T. (2015, May). An HMM-based speech-smile synthesis system: An approach for amusement synthesis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (Vol. 5, pp. 1-6). IEEE.

El Haddad, K., Torre, I., Gilmartin, E., Çakmak, H., Dupont, S., Dutoit, T., & Campbell, N. (2017, October). Introducing amus: The amused speech database. In *International Conference on Statistical Language and Speech Processing* (pp. 229-240). Springer, Cham.

Flom, R., & Bahrick, L. E. (2007). The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental Psychology, 43*(1), 238.

de Gelder, B., Böcker, K. B., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters, 260*(2), 133–136.

Green, J. R., Moore, C. A., & Reilly, K. J. (2002). The sequential development of jaw and lip control for speech. *Journal of Speech, Language, and Hearing Research, 45*(1), 66–79.

Heyes, C. M., & Ray, E. D. (2000). What is the significance of imitation in animals?. *Advances in the Study of Behavior, 29*, 215-245. Academic Press.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences, 5*(6), 253–261.

Heyes, C. (2016). Homo imitans? Seven reasons why imitation couldn't possibly be associative. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150069.

Heyes, C. (2021). Imitation. *Current Biology, 31*(5), R228–R232.

Heyes, C. (2023). Imitation and culture: What gives? *Mind & Language, 38*(1), 42–63.

Isomura, T., & Nakano, T. (2016). Automatic facial mimicry in response to dynamic emotional stimuli in five-month-old infants. *Proceedings of the Royal Society B: Biological Sciences, 283*(1844), 20161948.

Iyer, S. N., Denson, H., Lazar, N., & Oller, D. K. (2016). Volubility of the human infant: Effects of parental interaction (or lack of it). *Clinical Linguistics & Phonetics, 30*(6), 470–488.

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., … Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology, 26*(10), 1334–1338.

Jones, S. S. (2007). Imitation in infancy: The development of mimicry. *Psychological Science, 18*(7), 593–599.

Jones, S. (2017). Can newborn infants imitate? *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(1–2), e1410.

Kisilevsky, B. S., Hains, S. M., Lee, K., Xie, X., Huang, H., Ye, H. H., … Wang, Z. (2003). Effects of experience on fetal voice recognition. *Psychological Science, 14*(3), 220–224.

Kisilevsky, B. S., Hains, S. M., Brown, C. A., Lee, C. T., Cowperthwaite, B., Stutzman, S. S., … Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development, 32*(1), 59–71.

de Klerk, C. C., Lamy-Yang, I., & Southgate, V. (2019). The role of sensorimotor experience in the development of mimicry in infancy. *Developmental Science, 22*(3), e12771.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138–1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3), 361–381.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *The Journal of the Acoustical Society of America, 100*(4), 2425–2438.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science, 9*(2), F13–F21.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences, 105*(32), 11442–11445.

Lee, G. Y., & Kisilevsky, B. S. (2014). Fetuses respond to father's voice but prefer mother's voice after birth. *Developmental Psychobiology, 56*(1), 1–11.

Legerstee, M. (1991). The role of person and object in eliciting early imitation. *Journal of Experimental Child Psychology, 51*(3), 423–433.

Levitt, A. G., & Wang, Q. (1991). Evidence for language-specific rhythmic influences in the reduplicative babbling of French-and English-learning infants. *Language and Speech, 34*(3), 235–249.

Long, H. L., Bowman, D. D., Yoo, H., Burkhardt-Reed, M. M., Bene, E. R., & Oller, D. K. (2020). Social and endogenous infant vocalisations. *PloS One, 15*(8), Article e0224956.

Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current biology, 19*(23), 1994–1997.

Mastropieri, D., & Turkewitz, G. (1999). Prenatal experience and neonatal responsiveness to vocal expressions of emotion. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 35*(3), 204–214.

Maurer, D., & Lewis, T. L. (2001). Visual acuity: The role of visual input in inducing postnatal change. *Clinical Neuroscience Research, 1*(4), 239–247.

Messum, P., & Howard, I. S. (2015). Creating the cognitive form of phonological units: The speech sound 1correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. *Journal of Phonetics, 53*, 125–140.

McIntosh, D. N., Reichmann-Decker, A., Winkielman, P., & Wilbarger, J. L. (2006). When the social mirror breaks: Deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. *Developmental Science, 9*(3), 295–302.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*(4312), 75–78.

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

Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Infant and Child Development, 6*(3–4), 179–192.

Meltzoff, A. N., & Williamson, R. A. (2017). Imitation and modeling. *Reference Module in Neuroscience Biobehavioral Psychology*. https://doi.org/10.1016/B978-0-12-809324-5.05827-2

Meltzoff, A. N., Murray, L., Simpson, E., Heimann, M., Nagy, E., Nadel, J., ... & Ferrari, P. F. (2018). Re-examination of Oostenbroek et al.(2016): Evidence for neonatal imitation of tongue protrusion. *Developmental Science*, *21*(4), e12609.

Mersad, K., Kabdebon, C., & Dehaene-Lambertz, G. (2021). Explicit access to phonetic representations in 3-month-old infants. *Cognition, 213*, Article 104613.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development, 16*(4), 495–500.

Moran, G., Krupka, A., Tutton, A., & Symons, D. (1987). Patterns of maternal and infant imitation during play. *Infant Behavior and Development, 10*, 477–491. https://doi.org/10.1016/0163-6383 (87)90044-0

Morrish, K. A., Stone, M., Shawker, T. H., & Sonies, B. C. (1985). Distinguisability of tongue shape during vowel production. *Journal of Phonetics, 13*(2), 189–203.

Oller, D. K., Buder, E. H., Ramsdell, H. L., Warlaumont, A. S., Chorna, L., & Bakeman, R. (2013). Functional flexibility of infant vocalisation and the emergence of language. *Proceedings of the National Academy of Sciences, 110*(16), 6318–6323.

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C. C., … Vohr, B. (2019). Preterm and full term infant vocalization and the origin of language. *Scientific Reports, 9*(1), 14734.

Oller, D. K., Ramsay, G., Bene, E., Long, H. L., & Griebel, U. (2021). Protophones, the precursors to speech, dominate the human infant vocal landscape. *Philosophical Transactions of the Royal Society B, 376*(1836), 20200255.

Parga, J. J., Daland, R., Kesavan, K., Macey, P. M., Zeltzer, L., & Harper, R. M. (2018). A description of externally recorded womb sounds in human subjects during gestation. *PloS One, 13*(5), e0197045.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196.

Podesva, R., Callier, P., Voigt, R., & Jurafsky, D. (2015, August). The connection between smiling and GOAT fronting: Embodied affect in sociophonetic variation. In *ICPhS*.

Prinz, W. (2005). An ideomotor approach to imitation. *Perspectives on Imitation: From Neuroscience to Social Science, 1*, 141–156.

Quené, H., Semin, G. R., & Foroni, F. (2012). Audible smiles and frowns affect speech comprehension. *Speech Communication, 54*(7), 917–922.

Ray, E., & Heyes, C. (2011). Imitation in infancy: The wealth of the stimulus. *Developmental Science, 14*(1), 92–105.

Rayson, H., Bonaiuto, J. J., Ferrari, P. F., & Murray, L. (2017). Early maternal mirroring predicts infant motor system activation during facial expression observation. *Scientific Reports, 7*(1), 11738.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current Research and Theory, 2*, 64–69.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347–357.

Serkhane, J. E., Schwartz, J. L., Boë, L. J., Davis, B. L., & Matyear, C. L. (2007). Infants' vocalisations analyzed with an articulatory model: A preliminary report. *Journal of Phonetics, 35*(3), 321–340.

Slaughter, V. (2021). Do newborns have the ability to imitate? *Trends in Cognitive Sciences, 25*(5), 377–387.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics, 27*(1), 24–27.

Tartter, V. C., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *The Journal of the Acoustical Society of America, 96*(4), 2101–2107.

Titze, I. R. (1994). Fluctuations and perturbations in vocal output. *Principles of Voice Production*, 209–306.

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research : JSLHR, 50*(6), 1510–1545. https://doi.org/10.1044/1092-4388(2007/104)

Wass, S., Phillips, E., Smith, C., Fatimehin, E. O., & Goupil, L. (2022). Vocal communication is tied to interpersonal arousal coupling in caregiver-infant dyads. *ELife, 11*, e77399.

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology, 50*(1), 509–535.

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Current Directions in Psychological Science, 21*(4), 221–226. https://doi.org/10.1177/096372141244945

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science, 24*(5), 603–612.