Journal of Phonetics 41 (2013) 63–77

Contents lists available at SciVerse ScienceDirect

Journal of Phonetics

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

1765

Computational simulation of CV combination preferences in babbling

Hosung Nam [a], Louis M. Goldstein [a,b], Sara Giulivi [a,c], Andrea G. Levitt [a,d], D.H. Whalen [a,e,*]

[a] Haskins Laboratories, New Haven, CT, USA
[b] University of Southern California, Los Angeles, CT, USA
[c] DSLO-University of Bologna, Bologna, Italy
[d] Wellesley College, Wellesley, MA, USA
[e] City University of New York, New York, NY, USA

ARTICLE INFO

Article history:
Received 7 April 2011
Received in revised form
21 November 2012
Accepted 27 November 2012

ABSTRACT

There is a tendency for spoken consonant–vowel (CV) syllables, in babbling in particular, to show preferred combinations: labial consonants with central vowels, alveolars with front, and velars with back. This pattern was first described by MacNeilage and Davis, who found the evidence compatible with their ‘‘frame-then-content’’ (F/C) model. F/C postulates that CV syllables in babbling are produced with no control of the tongue (and therefore effectively random tongue positions) but systematic oscillation of the jaw. Articulatory Phonology (AP; Browman and Goldstein) predicts that CV preferences will depend on the degree of synergy of tongue movements for the C and V. We present computational modeling of both accounts using articulatory synthesis. Simulations found better correlations between patterns in babbling and the AP account than with the F/C model. These results indicate that the underlying assumptions of the F/C model are not supported and that the AP account provides a better and account with broader coverage by showing that articulatory synergies influence all CV syllables, not just the most common ones.

1. Introduction

The segmental inventories of languages differ, but there are strong universal tendencies (Maddieson, 1984). The stop manner of articulation is universal, and the three most common places of articulation are labial (99.1% of the 317 languages Maddieson surveyed; p. 32), alveolar (99.7%) and velar (99.4%). When children acquire language, their earliest productions have been described as having very similar syllables regardless of the child's language exposure (Locke, 1983; Oller, 2000), although there are language-specific effects as well (Boysson-Bardies, Hallé, Sagart, & Durand, 1989; Whalen, Levitt, & Wang, 1991). The universal consonant tendencies found in adult languages appear in most babbling inventories as well (Vihman, Ferguson, & Elbert, 1986). For vowels, languages tend to have /i/ (91.5%), /ɑ/ (88.0%) and /u/ (83.9%) (Maddieson, 1984, p. 125), with systems that range from three vowels to 24 (p. 126). The vowel systems distinctively use front, central and back places of articulation in 99.1% of the languages. These vowel qualities also occur in early babbling, with point vowels ([i, a, u]) being well-represented; these vowel are also among those mastered first in early word development (Stoel-Gammon & Herrington, 1990).

The combinations of these places of articulation of consonant and vowel, however, are not evenly distributed, a phenomenon first described by MacNeilage and Davis (1990a, 1990b). They discovered that certain consonant–vowel (CV) sequences occurred with greater than expected frequency as predicted by their ‘‘frame-then-content’’ account (hereafter, F/C) of language organization: Labial consonants with central vowels, alveolars with front, and velars with back. They found this both in babbling (Davis & MacNeilage, 1995, 1993; Zlatic, MacNeilage, Matyear, & Davis, 1997) and adult speech as estimated from dictionary entries (MacNeilage, Davis, Kinney, & Matyear, 2000).

Here, we will present simulations of the F/C account, contrasting it with an alternative based on the Articulatory Phonology (AP) of Browman and Goldstein (1986, 1989, 1992). AP takes articulatory gestures as the primitives of sound systems. They are organized into utterances via timing relationships, which can include overlapping of gestures. This account allows us to extrapolate from adult speech articulator compatibilities to what might be preferred in babbling, even if adult control parameters are not in place. We will then test both models with articulatory synthesis, employing a different synthesizer than that used previously to test the F/C account (Serkhane, Schwartz, Boë, Davis, & Matyear, 2007), one which avoids a confound between jaw motion and tongue motion.

[*] Corresponding author at: Haskins Laboratories, New Haven, CT, USA. Tel.: +1 203 865 6163; fax: +1 203 865 8963.
E-mail addresses: whalen@haskins.yale.edu, dwhalen@gc.cuny.edu (D.H. Whalen).

0095-4470/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2012.11.002

1.1. The frame-content account

All languages seem to have the syllable as a unit, and babbling is most easily heard as a sequence of syllables. Although it is clear that infants explore the potential phonetic space to a certain extent, there are also preferred productions that strike adult listeners as particular consonants and vowels. The F/C account posits that the earliest form of speech is the result not of speech per se but of rhythmic mandibular oscillation. The lips and tongue are hypothesized to not be actively controlled at this stage.

MacNeilage and Davis provided the following statistical analysis as the evidence for the jaw-only strategy. In utterances produced by infants beginning canonical babbling (approximately 7 months of age), they show a preference of a certain set of CV combinations: coronal consonants combine more readily with front vowels, labial consonants with central vowels, velar consonants with back vowels (Davis & MacNeilage, 1994, 1995; Davis, MacNeilage, & Matyear, 1999, 2002; Giulivi, 2007; Giulivi, Whalen, Goldstein, Nam, & Levitt, 2011). Tables 1 and 2 show the observed/expected ratios for each C–V combination averaged over all published results from MacNeilage and Davis's babblers (Davis & MacNeilage, 1994, 1995; Davis et al., 1999, 2002) and for the results for three language environments from Giulivi et al. (2011), respectively. The data are arranged so that all nine combinations of place (front, central and back for the vowels, labial, coronal and velar for the consonants) are represented. The diagonals of the tables are the combinations that were found to be preferred. No predictions were made for the off-diagonals. Ratios between observed and expected occurrence of syllables in those cells were calculated; those greater than 1 indicate that the combination is more frequent than one would expect by chance.

According to the F/C account, a CV-sounding babble is produced in the following two steps: (1) the vocal tract begins in a static (by hypothesis random) initial configuration and (2) the jaw's up and down movement generates CV-like sounds whenever the tongue or lower lip reaches the palate or the upper lip. What determines a consonant for a cycle of jaw movement is the location where the first constriction occurs. Vowels are determined by the initial vocal tract configuration and the degree of lowering of the mandible after the constriction. Further, the presetting of vocal tract before the jaw oscillation cycle is argued to tend to keep the initial shape by its inertia (Davis & MacNeilage, 2004). The mechanical jaw oscillation should then lead to a significant correlation between constriction location for consonants and vocal tract shape at the end of the cycle, which creates the co-occurrences of the CV components. Specifically, if the tongue happens to be advanced at the beginning of the mandibular oscillatory cycle, then the tongue tip will make contact with palate before the lips close, producing something that will sound like a coronal stop, and the lowered jaw position will sound like a front vowel. If the tongue is less advanced and low, the lips will contact each other before the tongue hits the palate, producing the percept of a labial stop, followed by a central vowel. If the tongue is raised and retracted, a perceived sequence of velar stop and high back vowel will result.

In this model, the lips and tongue are assumed to take on a fixed, but essentially random posture with respect to the jaw for the duration of the CV event. MacNeilage and Davis themselves do not refer to these positions as random, but their overt statements make it clear that only random positions make sense within the account. For example, they say that ''[m]ost of the variance in babbling, and early speech, is the result of mandibular oscillation, with other articulators tending to adopt a static configuration throughout the utterance'' (MacNeilage & Davis, 2000, p. 288). Davis is more explicit in later articles, such as the one with Serkhane and colleagues, stating ''at the onset of canonical babbling, the only active articulator is the lower jaw since the movements of the articulators it carries, that is, the lower lip and the tongue, as well as the velum and the upper lip, do not seem independent of the rhythmic jaw cycles'' (Serkhane et al., 2007, p. 322). The intent that the variety of syllable types will occur by chance positioning of the tongue is clear. Davis summarizes the approach this way: ''In this perspective, earliest vocal sequences are enabled by rhythmic jaw oscillations without independent movements of tongue or other active articulators from the onset of babbling (e.g., 'bababa'). … These jaw-cycle dependent patterns are predicted to differentiate into eventual 'content' or segmental elements (e.g. 'b' or 'd') as the child's production mechanism matures…'' (Davis, 2010, p. 306). Any deviation from randomness would imply control of the lips or tongue, but the ''lack of independent control of articulators other than the mandible during the basic oscillatory sequence of babbling'' (MacNeilage, 1998, p. 505) is foundational to the F/C account. If there were to be other processes at work, they have yet to be described, and it is difficult to see how control and lack of control could be reconciled. Thus we tested the only version of the F/C account that has been made moderately explicit: random initial states of the articulators.

Although this account sounds plausible, there are some problems with it. First, these CV combination biases are also found, somewhat less extremely, in adult lexicons. Adults do control their tongue and lips, so there must be some other principle than the F/C account at work in producing these patterns (Goldstein, Byrd, & Saltzman, 2006). Further, the diagonals (preferred combinations) account for only about half of the total productions in babbling, suggesting either that some aspect of the jaw motion results in non-preferred combinations or that other articulators are being manipulated, or at least moving during the duration of the CV, even at the earliest stage (Giulivi et al., 2011). These off-diagonals have not been explored for any systematicity. If there is a pattern in the less preferred combinations as well as in the preferred ones, then there needs

Table 1
Babblers' CV probability ratios from MacNeilage and Davis.

	Front	Central	Back
Coronal	1.38	0.69	0.94
Labial	0.68	1.31	1.02
Velar	1.05	0.87	1.14

Table 2
Babblers' CV probability ratios from Giulivi et al. (2011).

	Front	Central	Back
Coronal	1.07	0.80	0.85
Labial	0.82	1.53	1.32
Velar	1.05	0.89	0.99

to be a model that can account for both the diagonals and non-diagonals. Finally, kinematic analysis of infant feeding indicates ability to actively coordinate activities of the lips, tongue and jaw (see discussion in Whalen, Giulivi, Goldstein, Nam, & Levitt, 2011), rather than the lack of control of the lips and tongue assumed by F/C. What is needed is a second model that also assumes a biomechanical origin for these combination preferences but does not depend on the same parameters as the F/C model.

1.2. Articulatory Phonology

An alternative account based on Articulatory Phonology (AP) relates the CV biases observed in infant utterances are due to inherent compatibilities or synergies in the speech production system. In AP, the vocal tract is decomposed into distinct constricting devices (lips, tongue tip, tongue body, velum and glottis), which are used to form linguistic gestures, i.e., constriction tasks at vocal tract locations appropriate to those constrictors. Gestures for oral constrictors (lips, tongue tip, tongue body) are defined by constriction location and degree task variables: lip protrusion (LP)/lip aperture (LA), tongue tip constriction location (TTCL)/tongue tip constriction degree (TTCD), and tongue body constriction location (TBCL)/tongue body constriction degree (TBCD). Since the location of non-oral constrictors (velum and glottis) does not vary, they only use constriction degree variable: VEL (velum) and GLO (glottis). Thus, these eight ‘‘vocal tract variables’’ constitute the dimensions along which constricting tasks are defined. Importantly, control of each tract variable harnesses a set of articulator degrees of freedom that can contribute to the ongoing constricting action into a coordinative structure (Saltzman & Munhall, 1989). Individual articulator degrees of freedom can participate in more than one tract variable coordinative structure. Table 3 shows the set of tract variables and their associated articulator model degrees of freedom. For example, tongue body articulators (CL, CA) are shared by tongue body tract variables (TBCL, TBCD) and tongue tip tract variables (TTCL, TTCD). The jaw articulator (JA) is shared by lip tract variables (PRO, LA), tongue body tract variables (TBCL, TBCD) and tongue tip tract variables (TTCL, TTCD). See Section 3.1 for more detailed geometric description of the articulators.

A word can be described as a constellation of gestures, and the lexical contrast between words can be related to the presence or absence of gestures and the relative timing among them, all of which are encoded for each word in a *coupling graph* (Goldstein et al., 2006). Each gesture is specified with its constriction location and degree information. For example, /s/ involves a tongue tip constriction gesture, which is defined with a ‘critical’ (for the production of turbulence) constriction degree at ‘alveolar’ constriction location. /p/ and /n/ involve complete constriction (‘closed’ constriction degree), one employing the lip constrictor and the other the tongue tip constrictor at the ‘alveolar’ constriction location. The velum and glottis gestures are defined only by their constriction degree (‘closed’ or ‘wide’) because their constrictions occur at fixed locations.

From the perspective of AP, the child is thought to inherit some control of the lips, tongue tip, and tongue dorsum constriction devices, as well as the jaw from ancient mammalian capabilities (Studdert-Kennedy & Goldstein, 2003). Each of these systems is considered an ‘‘organ’’ capable of being controlled to a certain extent. They suggest that this proposal accounts for the early emergence of between-organ distinctions in the child’s production of early words ([b] vs. [d], for example) compared to the late emergence of within-organ distinctions ([b] vs. [f], for example). Thus while assuming some level of control of articulators other than the jaw strikes some researchers as unparsimonious, it does do a better job of accounting for a wide range of findings. It is not, strictly speaking, an evolutionary account, but it makes the assumption that there was a chain of adaptations of facial imitation (common to hominids) to vocal imitation (apparently unique to humans among primates) (Studdert-Kennedy & Goldstein, 2003, p. 247).

Here we hypothesize that the infant’s early babbling involves oscillatory action (closing and opening) of one (or more) of the constricting organs: lips, tongue front (tip), tongue rear (body), which is in part an imitative response to the actions of those organs perceived by the infant using visual and/or auditory input. The closed phase of the oscillation will be perceived by an adult as a consonant (stop or glide), whose place of articulation depends on which organ is oscillated, and the open phase as some vowel. At this stage, there is no assumption that this oscillatory control is decomposed temporally into a shorter consonant constriction gesture and a longer vowel gesture. The closed–open properties of babbles emerge from oscillations, as in the F/C model, but those oscillations are not limited to the jaw, and they can involve synchronous oscillation of a consonant-related organ and a vowel-related organ. Further, oscillation of a constriction organ is hypothesized to involve active engagement of multiple articulatory components (somewhat like the coordinative structures that underlie adult gestures). Infants certainly come into the world actively coordinating their lips and tongue along with the jaw to produce feeding behavior (see discussion in Whalen et al., 2011). So for example, oscillating the lips between open and closed is hypothesized to engage the jaw as well the lips (just as it does, for example in monkey lip-smacks, e.g. Redican, 1975). Crucially, this jaw activity will also result in passive displacement of the tongue with respect to the palate. Thus, the infant’s oscillation of the lips (including supplemental motion of the jaw) will tend to (passively) produce certain patterns of tongue body constriction, and adults will hear these tongue body constrictions as instances of particular vowels. These are the vowels whose tongue body constrictions can be described as being highly *synergistic* with the lip action: their tongue body constrictions will tend to be produced by a lip oscillation, regardless of whether or nor the tongue body constriction itself is actively controlled. A different set of vowel-like tongue body constrictions will be maximally synergistic if the infant oscillates the front part of the tongue (or tip) and yet another set if the infant oscillates the rear of the tongue. The relative likelihood of babbled syllables being perceived as a particular consonant–vowel combination is thus

Table 3
Vocal tract constrictors, variables and associated articulators and their variable names in parentheses (see Fig. 1 and Section 3.1 for more details and an explanation of the labels associated with the articulators).

Constrictors	Tract variables	Articulators involved
Lips	Lip protrusion (PRO) Lip aperture (LA)	Upper lip (UH), lower lip (LH), jaw (JA) Upper lip (UH), lower lip (LH), jaw (JA)
Tongue tip	Tongue tip constriction location (TTCL) Tongue tip constriction degree (TTCD)	Tongue tip (TL, TA), tongue body (CL, CA), jaw (JA) Tongue tip (TL, TA), tongue body (CL, CA), jaw (JA)
Tongue body	Tongue body constriction location (TBCL) Tongue body constriction degree (TBCD)	Tongue body (CL, CA), jaw (JA) Tongue body (CL, CA), jaw (JA)
Velum	Velum (VEL)	Velum (NA)
Glottis	Glottis (GLO)	Glottis (GW)

hypothesized to result from the degree of synergy between a given oscillating organ (corresponding to the consonant) and the tongue body constriction associated with that type of vowel.

Adult dictionary counts also show the same CV preferences (MacNeilage et al., 2000; Whalen et al., 2012), but not all languages show the same preference in spoken corpora (Whalen et al., 2012). This suggests that adults, despite having greater control over separate gestures for consonants and vowels, nonetheless exhibit an influence of the same synergies when it comes to selecting new lexical items. In order to understand this preference, we need to consider how consonants and vowels are combined to form CV syllables. Phoneticians have long noted differences in how consonants pattern at the beginning (onset) and end (coda) of syllables. Recently, gestural studies of syllable structure (Goldstein et al., 2006; Nam, 2007; Nam & Saltzman, 2003) have found evidence that such positional differences (onset vs. coda) in syllable structure can be understood as a difference in mode of coordination between C and V gestures: timed so that their activations begin at the same time (‘‘in-phase’’) for CV or timed so that one begins half-way through the other (‘‘anti-phase’’) for VC (Browman & Goldstein, 2000; Goldstein et al., 2006; Nam, 2007). Simulation and experimental studies (Mooshammer et al., 2012; Nam, 2007; Nam, Goldstein, & Saltzman, 2009; Nam & Saltzman, 2003) have further shown that the distinct modes can predict such observed patterns as: (1) greater inter-gestural timing variability in coda clusters than in onset clusters; (2) faster reaction time to initiate production of CV rather than VC words; and (3) earlier emergence of CV syllables than VC syllables in infants, but also (somewhat paradoxically) earlier emergence of consonant clusters in coda than in onset.

Goldstein et al. (2006) hypothesized that if CV combinations are produced with synchronous (in-phase) triggering of C and V gestures, then those combinations of C of V that exhibit greater synergy (in the sense defined above) would be preferred over others with less synergy, because all the articulatory motions for both C and V can readily be produced synchronously, at least in adults. This follows from the fact that the articulatory motions for the C and the V in synergistic C–V pairs are largely shared: the tongue body movements required for the V would be produced passively by the movements required for the C. This account predicts synergy-related CV preferences but not VC ones in adults because, according to this hypothesis, the latter are coordinated in anti-phase mode, for which synergy is not relevant. Consistent with this view, systematic combination preferences, like those found with CVs, have not been reported for VCs (MacNeilage et al., 2000). In children, however, MacNeilage et al. (2000) did find such preferences. In the account we are pursuing here, these preferences, like the CV ones, result from continuous oscillations of the organ (corresponding to the consonant) and the tongue body constriction associated with that type of vowel, which would be assumed to be in-phase oscillations. The factors that cause individual babbled syllables in other than final position to perceived (transcribed) as VC, as opposed to CV, are unknown, but could include how phonation (or lack of it) reveals particular chunks (phases) of the oscillatory cycle.

1.3. Means of evaluating the hypotheses

The preference for certain CV combinations has been found repeatedly, but there is currently no study directly measuring the articulators during babbling in order to judge which biomechanical model might be supported. Such data are difficult to obtain, although ultrasound measures are showing promise. Instead, we can make use of articulatory synthesis, in the form of a software synthesizer that allows independent control of tongue, jaw and lips, to model the systematic variation associated with the F/C and the AP accounts.

2. Computational simulations

2.1. F/C account

Two simulation studies have attempted to evaluate the F/C account (Serkhane et al., 2007; Vilain, Abry, Badin, & Brosda, 1999) using articulatory synthesis. The studies used either Maeda's model (Maeda, 1990) or a variant of it. Vilain et al. (1999) claimed general support for the F/C account, but their report does not include a detailed exploration of the babbling articulatory space. Serkhane et al. (2007) modeled babbling by inferring initial articulatory configurations from 7-month-old vocalizations. Then, they varied the ‘‘jaw’’ parameter to create an upward oscillation from these inferred vocal tract shapes until contact was made somewhere along the vocal tract. The combination of consonant-like sounds at the closure and the vocalization was considered a CV sequence. The majority of their front vowel syllables contained an alveolar consonant, as expected from the babbling data. However, the majority of syllables with central vowels also contained alveolar consonants rather than the expected labial (which accounted for only 22% of the syllables). Labials dominated with the back vowels, rather than the expected velars. Thus only one of the three vowel places of articulation resulted in the expected pattern.

In addition to the overall lack of correspondence between the model and the data, there is an important limitation in the interpretation of these results: The manipulations of the ‘‘jaw’’ component necessarily entail tongue shape changes. Maeda (1990) modeled cineradiographic images and labiographic data of speech and began by factoring out jaw motion with Principal Component Analysis (PCA). Then, the PCA was performed on the residual to find the contribution of the tongue and lips. Thus the contribution of jaw movement to tongue shapes through linear regression might not correspond to the purely *anatomical/biomechanical* contribution of the jaw to the tongue shape, which is what is required to test the F/C account (see also de Boer & Fitch, 2010). Specifically, French (on which the model was based) has many items with the vowels /i/ and /ɑ/ in its vocabulary. We might expect jaw and the overall tongue body to be raised together for /i/ and lowered for /a/ or /ɑ/ (cf. the English data of Whalen, Kang, Magen, Fulbright, & Gore, 1999). There should be a high correlation between jaw and tongue's vertical position. But because /i/ also involves a substantial amount of tongue root advancement in French and English, the advancement will also be highly correlated with jaw movement even though such advancement is not in fact caused anatomically or biomechanically by the jaw movement, as can be seen, for example, in Advanced Tongue Root vowels in Akan (Tiede, 1996).

In this study, we employed the Configurable Articulatory SYnthesis model (CASY: Rubin et al., 1996) developed at Haskins Laboratories. It avoids the confound of Serkhane et al.'s simulation, because the jaw control parameter affects only the position of the tongue relative to the hard palate, not its shape. As shown in Fig. 1, the position of the tongue body (C) is defined as CA and CL on the jaw coordinate (FJ) as is determined by anatomy and geometry. Throughout the simulations, we fixed the tongue body articulator variables, CL and CA to simulate pure jaw oscillation. That is, as the jaw raises and lowers with the tongue body position fixed, the tongue body will be passively carried along with it. CASY is based on Mermelstein's articulatory model (Mermelstein, 1973), where major articulators (jaw, tongue body, tongue tip, etc.) are parametrically

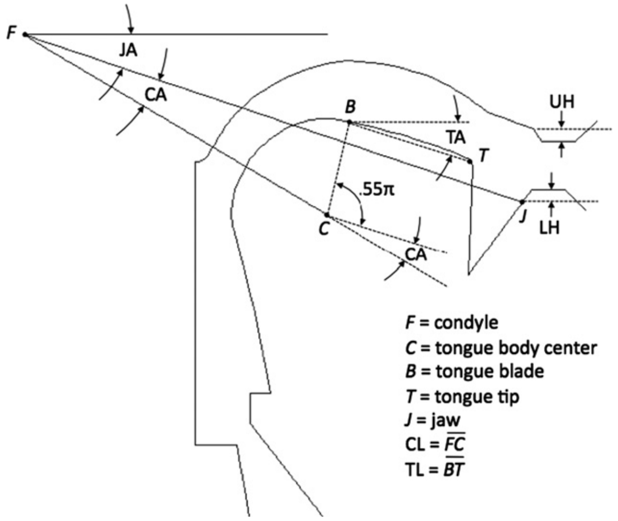

Fig. 1. Vocal tract representation and model articulator variables of CASY: upper lip (UH), lower lip (LH), tongue body (CL, CA), tongue tip (TA, TL), jaw (JA). The position of jaw (J) is given with respect to the mandibular condyle (F). Since the distance of the jaw from the condyle is not variable, the jaw position can be given by just the angle (JA) from a horizontal line passing through the joint. The tongue has a tongue body component and a tongue blade/tip component. The tongue body is represented as a circle of a fixed radius with the movable center (C). The tongue tip (T) and blade are attached onto it. The position of tongue-body circle center (C) is given as the angle (CA) with reference to condyle-to-jaw line and the length (CL) of C from the condyle bone (F). The tongue blade is attached to the tongue body circle at 0.55π–JA (B) with respect to the horizontal line passing through C. The position of the tongue tip (T) is defined as B-to-T angle (TA) with respect to a horizontal line passing through B and the length (TL) of T from B. In sum, Jaw (J) is represented as JA, tongue body (C) as CL and CA with respect to jaw (J), and tongue tip (T) as TL and TA with respect to tongue body.

controlled, but with the extension that the overall length and shape of the hard structures of the vocal tract can be adjusted to different speakers, including infants.

Simulations of the F/C account were conducted for both the adult and 7-month-old infant vocal tracts. We used a default vocal tract configuration of CASY for the adult simulation. For the infant simulation, we adjusted the vocal tract shape to a 7-month-old one using MRI images and measurements from Vorperian (2000) and Vorperian et al. (2005). For both simulations, we generated random initial vocal tract configurations by varying relevant articulator parameters and created mandibular raising and lowering by controlling the jaw parameter.

2.2. Articulatory Phonology

A second pair of computational simulations was designed to examine AP predictions that articulatory synergy in different CV combinations would model the babbling data. For this, we used the TAsk-Dynamic Application model (TADA; Nam, Goldstein, Saltzman, & Byrd, 2004). TADA is a computational speech production model in which AP is mathematically implemented. An utterance is represented as an ensemble of constricting actions, or gestures, of five distinct constrictors (lips, tongue tip, tongue body, velum, and glottis) in the form of a gestural score. Each gesture is associated with a relevant set of model articulators (see Table 3), which cooperate to achieve the gestural target. For example, lip closing for /p/ is defined as a closure target for lip aperture and engages the following articulator degrees of freedom: UH (upper lip vertical displacement (“height”)), LH (lower lip vertical displacement), and JA (jaw angle) articulators. The gestural target of closing lips will always be achieved, but the articulators’ cooperative movements will be determined in contextually distinct ways. In the model, a dynamical control regime defined in the task- (constriction-) coordinate space, whose parameters are time-invariant (throughout the interval during which a gesture is active), is transformed to time-varying, posturally-dependent dynamical parameters in the model articulator coordinate space. The transformation is accomplished through a weighted pseudo-inverse of the Jacobian matrix that relates change in articulator state to change in constriction variable state (Saltzman & Munhall, 1989; Whitney, 1969). The weighting parameters (or articulator weights) modulate how much each articulator will be engaged by that constriction variable, everything else being equal. The task-dynamic procedure computes each tract variable’s trajectory and coordinated movement trajectories of its associated articulators. CASY, which is incorporated into TADA, then estimates corresponding area functions from the computed articulator states at a given time point and generates acoustic signals.

In the current simulations, TADA was used to model C and V gestures separately and to compare the output tongue body positions (tongue body center, or “TBC”) for each, in order to determine the degree of synergy of C–V gesture pairs. If a C and V are synergistic with respect to the tongue body (Fig. 2a), they will have final positions that are closer to each other than for less synergistic pairs (Fig. 2b). Specifically, the Euclidean distance between the output TBC positions can be assessed for any pair of C and V. The TBC output difference between C and V can be used to calculate the amount of synergy that would be exhibited in adult speech as scaled down to an infant’s vocal tract size. This does not imply that these adult targets were also the targets of the babbling; rather, they are simply a way of evaluating similarity between the tongue body constriction produced passively during a consonant-type oscillation of the lips, tongue front (tip), or tongue dorsum and the tongue body constrictions for front, central and back vowels. TADA is not being used here to produce simulated babbling output that would be perceived as CV syllables. Indeed, that is not how the output would be perceived. The temporal control hypothesized for babbling would be oscillatory, not the temporal control of adult gestures. The point of the TADA simulation is simply to reveal the consequences of making different consonant-type constrictions (in terms of the articulatory coordination and magnitude, not timing) on the tongue body, and to see how similar these consequences are to particular vowels.

3. Simulations

For the F/C simulation, we manipulated CASY only (not the TADA model) because the articulator variables in CASY can be explicitly controlled according to the factors specified in the F/C account. On the other hand, we employed TADA for the AP simulation because TADA

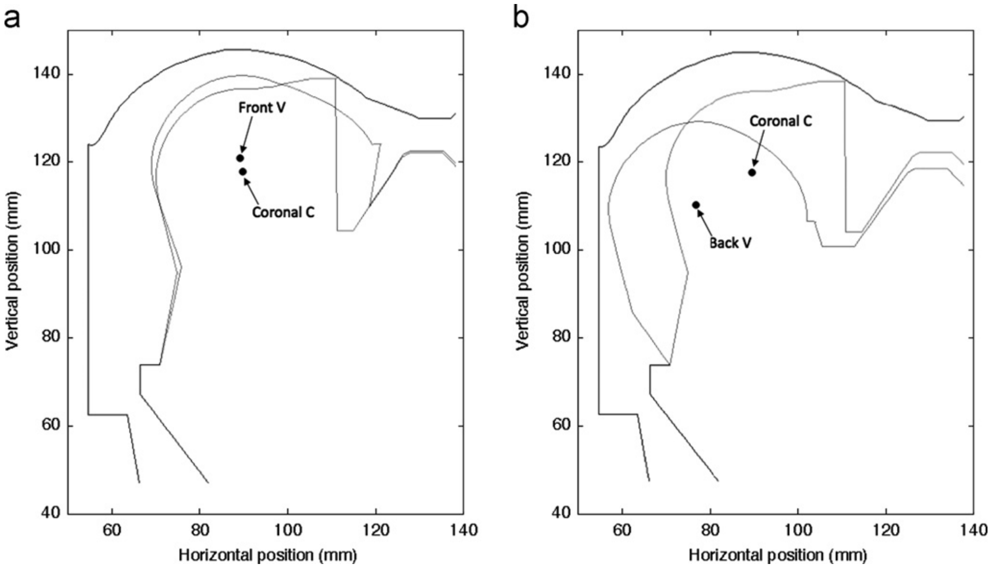

Fig. 2. Superimposed final vocal tract shapes and tongue body center positions (a) when C and V are synergistic and (b) when C and V are not synergistic: (a) a coronal C (TTCL 66 degrees, TTCD 0 mm) and a front V (TBCL 84 degrees, TBCD, 6 mm), (b) a coronal C (TTCL 66 degrees, TTCD 0 mm) and a back V (TBCL 176 degrees, TBCD, 6 mm). The origin point is an arbitrary location used within CASY.

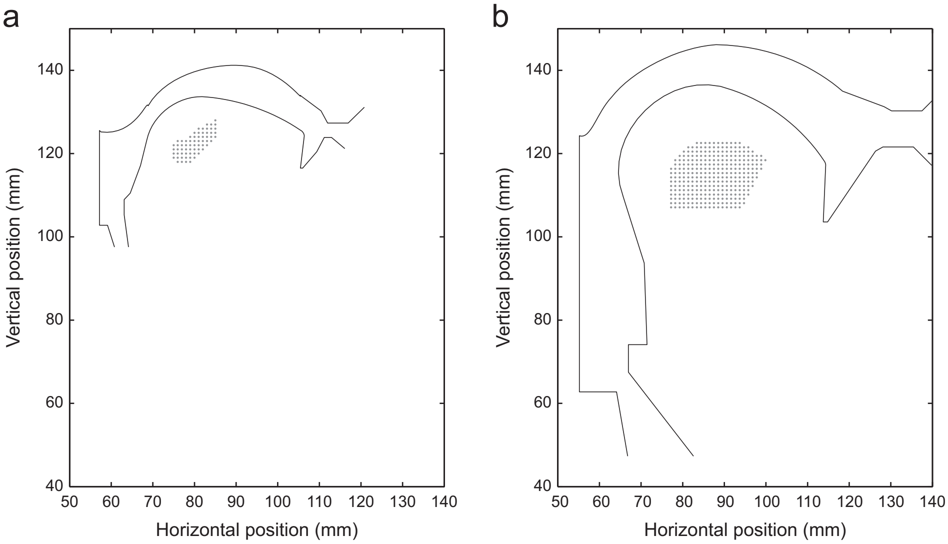

Fig. 3. CASY vocal tract configurations of (a) 7-month-old infant and (b) adult. Tongue body center (black dot = neutral position) and its physiologically plausible space (gray dots) for vowels. (The origin point of each of the two axes is an arbitrary location used within CASY.)

allows us to model the inter-articulator coordination patterns that might plausibly underlie oscillating lip, tongue tip and tongue body constrictions, and then to evaluate the degree of tongue body synergy between those C-type constrictions and particular Vs, as necessary for testing that account.

3.1. Simulation 1: The F/C account

3.1.1. Method

CASY was used to model syllables articulated solely by movement of the mandible. In CASY, the positions of all the articulators (jaw, tongue body, tongue tip, lips) are represented as polar coordinates (distance and angle) with respect to reference sites or other articulators (see Fig. 1).

Simulations of the F/C account were carried out for both adult and 7-month-old infant vocal tract configurations. The default configuration of CASY, which is based on sagittal x-ray images of adult vocal tracts, was used for the adult vocal tract (Fig. 3b). The CASY vocal tract configuration for babblers (Fig. 3a) was estimated by visually fitting the CASY vocal tract outline to that of a 7-month-old infant's midsagittal MRI image (Vorperian et al., 2005) using a visual fitting tool. Although this gives good midsagittal fits, the effect on the distance-to-area functions necessary for generating three-dimensional tubes is unknown. Here, we make the assumption that the cross-sectional area for a constriction of a given distance will be similar to that of adults, but it could well be that three-dimensional imaging of infant vocal tracts would reveal differences.

According to the F/C account, a cycle of oscillation starting from an arbitrary initial vocal shape results in a CV-like sound. Random configurations of the initial vocal tract shape were obtained by varying five articulator variables in CASY: JA, CL, CA, TL, and TA (see Fig. 1). Reasonable ranges of the variables were chosen to cover physiologically plausible configurations, without making an initial constriction along the vocal tract. Once a random vocal tract shape was selected, a cycle of up-and-down jaw (J) movement was modeled by varying the jaw angle (JA). As the jaw moved up and down, the other variables, CA, CL, TA, and TL were kept constant. The jaw was raised until either the tongue or the lower lip made contact with a fixed structure, creating a constriction. The constriction site was identified by finding the section where the area function was zero along the vocal tract, with the possible regions being labial, coronal or velar. Due to the nature of the articulators and their movement, there were clear clusters at the coronal and velar places, with no ambiguous locations. Once an obstruction in the vocal tract occurred, the jaw was lowered by modulating JA. How far the jaw was lowered, that is, the JA value at the landing site of jaw lowering, was randomly selected within a physiologically plausible range. During a complete cycle for producing a CV (starting with a random vocal tract shape and then adding jaw raising and jaw lowering), cross-sectional area functions were estimated at each movement frame and resonance characteristics were then computed to generate acoustic signals. Any physiologically implausible configuration before or after jaw oscillation was excluded. The simulation was continued until 1000 trials were obtained. We employed tongue body position (more precisely, tongue-body circle center position (C in Fig. 1)) in CASY to identify vowel type as front, center, or back.

In order to classify the resulting tongue body positions as front, center or back, we created a reference set of 300 physiological plausible tongue body center positions for the adult vocal tract, and 65 positions for the child (see Fig. 3), and then divided the resulting two-dimensional space of points into regions corresponding to front, central, and back vowels. The region boundaries of this map were then used to classify final tongue body center position of each F/C simulation. The partitioning of the space of the vowels into regions was done using three separate criteria: articulatory, acoustic, and perceptual. In computing the acoustics (and therefore subsequent perception) of the vocal tract shapes resulting from the tongue body positions, all other articulators, including the jaw, were fixed. Even though jaw height will affect vowel identity, the fixed jaw assumption is reasonable, as we are only attempting to classify the front–back classification of the tongue body positions, not the degree of openness (which is what will vary as a function of jaw height).

For the adult map, 41 of the 300 CASY-generated vowels for the adult tract were used for the perception test. This selection included some of the extreme values but concentrated on the regions around the expected boundaries. For the infant map, and all 65 of the vowels were used for the perception test. Each of these tongue body center positions is associated with a set of formants that is generated by the model. We used the formant information (F1 and F2) to identify the boundaries for front, central, and back vowels as follows. For the adult tract, tongue body center positions with F2>1800 Hz are regarded as front, those with F2<1800 Hz and F1<550 Hz as central, and those with F2<1800 Hz and F1>550 Hz as back. Note that these are not the same values that we would assume from measurements of natural productions because there was no secondary lip rounding associated with the back vowels. Second, we performed a perceptual test with three phoneticians. We presented them with the 41 vowels for the adult tract and the 65 vowels for the infant tract for judgment of place as front, central, or back. Their judgments were converted to 1, 2 and 3 for front, central and back, averaged and used to find perceptual boundaries as follows: front (avg<1.5), central (1.5<avg<2.5), and back (avg>2.5). The inter-rater agreements were 68% for the adult speech and 72% for the infant. Third, we examined the area functions of the 300 or 65 vowels and identified the location exhibiting the minimal constriction along the vocal tract. The minimal constriction locations did not always show gradual change but rather exhibited discrete shifts in constriction location (cf. Iskarous, Nam, & Whalen, 2010b). We estimated the constriction location boundaries by using the discrete constriction location boundaries where they occurred supplemented by the acoustic boundaries.

The articulatory boundaries for the child vocal tract did not match the acoustic and perceptual boundaries as well as those of the adult vocal tract. This highlights the limitations that we can expect from applying simple scaling to a tongue model that is based on adult data (de Boer & Fitch, 2010). However, there is insufficient information about child articulation to create a more accurate model. There is modeling evidence that suggests that infants may use different articulatory shapes to achieve relatively adult-like vowel percepts (Ménard, Davis, Boë, & Roy, 2009; Ménard, Schwartz, & Boë, 2004), but the results are insufficient for implementing a new production regime for infant vocal tracts. We will have to wait for better articulatory data before we can be sure that we have accurately represented the differences. In the mean time, the fact that the infant vocal tracts in the Ménard et al. studies elicited vowel percepts throughout the vowel space indicates that even the limited modeling that we have at our disposal accurately generates the kinds of patterns that are found in babbling transcriptions.

3.1.2. Results

In the jaw-oscillating (F/C) simulation, we measured tongue body position at the last frame of the jaw oscillation and compared it to the three types of vowel maps for both the adult simulation (Fig. 4) and the infant simulation (Fig. 5) to identify vowel type. The consonant classification

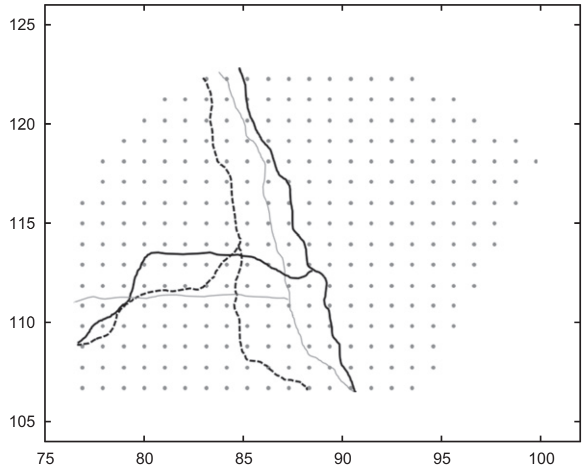

Fig. 4. Articulatory (gray), acoustic (black) and perceptual (dotted) maps of tongue body position for three vowel types (front, central and back), superimposed on the vowel space generated in Fig. 3, adult vocal tract.

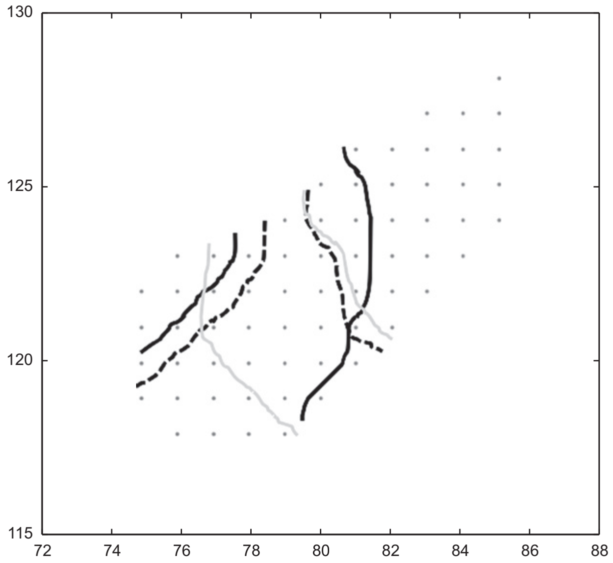

Fig. 5. Articulatory (gray), acoustic (black) and perceptual (dotted) maps of tongue body position for three vowel types (front, central and back), infant vocal tract.

Table 4
CV probability ratios from F/C simulation (7-month-old infant) for acoustic/articulatory/perceptual maps.

	Front	Central	Back
Acoustic			
Coronal	1.09	1.01	0
Labial	1.56	0.97	0.0
Velar	0.12	1.05	2.94
Articulatory			
Coronal	1.09	2.27	0.49
Labial	1.56	0.96	0.94
Velar	0.12	0.51	1.31
Perceptual			
Coronal	1.24	1.03	0.30
Labial	1.53	0.99	0.25
Velar	0.09	0.99	2.43

was determined by noting whether it was the lower lip, tongue tip, or tongue body that produced the closure on a given simulation.

We counted the number of occurrences for each CV combination and normalized the ratio using the same formula as Davis & MacNeilage (1995) and Giulivi et al. (2011) as in Tables 1 and 2.

Normalized CV ratio = P(CV)/(P(C) × P(V))
P(CV) = Sum of occurrences of a given CV/Total sum of occurrences
P(C) = Sum of occurrences of a given C/Total sum of occurrences
P(V) = Sum of occurrences of a given V/Total sum of occurrences

The results of the acoustic, articulatory, and perceptual maps give qualitatively similar results, except for the coronal+front combination in the adult simulations. Since analysis of babbling corpora are based on transcribed phonetic categories, we focus on the perceptual classifications. Adult F/C simulation results (Table 5) show ratios greater than one for all three diagonals. For the infant simulations (Table 4), the labial-central value is just less than one (0.99), and perhaps more importantly, it is not nearly as high as the labial-front combination (1.53). Further, the ratio for the velar-back is much larger than that found in natural corpora (Davis & MacNeilage, 1994, 1995; Davis, MacNeilage, & Matyear, 1999, 2002; Giulivi et al., 2011). Thus, the pattern of ratios on the diagonals that emerges in the F/C simulation provides at best limited support for the F/C account. For off-diagonals, the F/C simulation exhibits ratios for velar-front (for infants) and coronal-back (for both adults and infants) that are much lower than those reported from babbling. The absence of any labial-back combinations, however, is reminiscent of the data for hearing impaired children (von Hapsburg, Davis, & MacNeilage, 2008). A detailed analysis with quantitative comparisons to observed ratios will be conducted in Section 4.

3.2. Simulation 2: Articulatory Phonology

3.2.1. Method

For our simulation of the AP account, we employed the TADA parameterization of consonants and vowels in order to determine the relative synergy of pairs of consonant and vowel gestures. For adults we hypothesized that the maximally synergistic pairs should be easiest to

Table 5
CV probability ratios from F/C simulation (adult) for acoustic/articulatory/perceptual maps.

	Front	Central	Back
Acoustic			
Coronal	0.18	1.17	0.08
Labial	0.97	1.11	0.92
Velar	1.91	0.44	2.18
Articulatory			
Coronal	0.39	1.20	0.33
Labial	0.98	1.11	0.93
Velar	1.6	0.28	1.90
Perceptual			
Coronal	1.02	0.96	0.07
Labial	1.01	1.17	0.89
Velar	0.93	0.43	2.28

synchronize in CV syllables, and therefore should occur most frequently. For infants, we did not assume that the model's targets would be the specific target of any infant production. Rather, we used the tongue-body synergy measure between individually produced vowel and consonant gestures to determine what vowel positions would most likely emerge as passive consequences of producing oscillatory consonant-type gestures of lip, tongue tip, and tongue dorsum.

Consonants at a single labial location and a range of coronal and velar locations were modeled as lip, tongue tip, and tongue body gestures, respectively. Vowels were modeled as tongue body gestures, whose constriction location and degree targets were chosen to correspond to the 300 systematically varying locations of tongue body center within the vocal tract (Fig. 3). Articulatory trajectories for the formation of bare consonants (that is, movement from a neutral configuration to consonant closure) and bare vowels (movement from a neutral configuration to vowel target) were computed using the task-dynamic model. One possible way to measure the synergy is to compare movement vectors (with direction and distance) between two configurations. Indeed, subtracting one vector from another provides a measure of how different two vectors are. However, vector subtraction is problematic in our case because it depends on the starting position of the vocal tract. Fixing the starting position (e.g. the position for schwa) could constrain the variation but the simulation result would then be biased because articulatory movements in real speech would not always start from that position. Instead, the Euclidean distance between the end positions of the tongue for the two positions (C and V) is the most direct measure of articulatory compatibility for these syllables. We thus had our model produce bare consonant and vowel trajectories and used their final configuration for measurement. Tongue body center (TBC) position is the major end effector for vowel constrictions, but it necessarily takes on a value for each consonant. How close the TBCs for a given pair of vowel and consonant are when they reach their targets was our measure of how synergistic the two gestures would be if they overlap in time (Fig. 2). This can be related to the notion of tongue-body synergy which Iskarous, Fowler, and Whalen (2010a) used to show that the CV locus equations arise due to the synergy between the C and V.

To ensure that both the vowel and consonant datasets were sufficiently random such that simulation results are not attributed to articulatory presets from a particular language, randomness was provided in two ways. In the first simulation, the gestures were modulated by the constriction location targets so that they were not those of a specific language. In the second simulation, a single location for place of articulation for each of the three consonant places was taken as the starting point but the articulator weights varied randomly. As described in Table 3, a gesture is associated with a set of articulators that jointly contribute to achieving the gesture's target. For example, tongue tip gestures reach their targets by the cooperative engagement of the associated articulators, jaw, tongue body, and tongue tip. Such articulators' contributions can be asymmetrically implemented by assigning different weights to them such that if an articulator is ''lighter'' than the others, it is more influential to the gestural realization. Heavier articulators are harder to move, and therefore influence the vocal tract shape less than lighter ones. Since the articulator weights can differ from speaker to speaker and even from language to language, and since infants cannot be presumed to have any specific pattern of weights, it was necessary to vary the weights, as these would be expected to influence the measured C–V synergy.

The vowel dataset generated in Section 3.1 was used in the AP simulations. Hence, the target parameters of TBCL and TBCD gestures for vowels corresponding to each TBC position do not represent a vowel space for a specific language but rather a systematic sampling of typical vowel spaces. The vowel gestures, TBCL and TBCD, are realized by jaw (JA) and tongue body (CL, CA) articulators. We set all the articulator weights to the same value to exclude language-specific articulator synergy settings. The vowel dataset was compared to the consonant datasets which were modulated by constriction location and articulator weight, as detailed in Sections 3.2.1 and 3.2.2, respectively. Because the task-dynamical control model in TADA currently functions only for an adult vocal tract, the simulations were only carried out on the adult vocal tract.

3.2.1.1. Constriction location modulation. For initial consonant targets, we generated a range of gestures for each consonant type: lip (LA, PRO), tongue tip (TTCL, TTCD), and tongue body (TBCL, TBCD) gestures for labial, coronal, and velar stops, respectively. Their constriction locations were modulated as follows. Location for labials was not varied because physiologically plausible differences in the location of a labial constriction are small. Five locations in the coronal region, using the tongue tip, and five in the velar region, using the tongue body, were used to simulate variability in consonant production. Note that constriction degree was fixed to 0 for all the consonant types to ensure complete constriction. Constriction location values were specified using the angle of a polar grid, in which 0 degrees represents the location of the upper lip, 90 degrees represents the center of the hard palate, and 180 degrees represents the mid-pharynx. The tongue tip closure location varied from 46 to 86 degrees (alveolar to palatal) and the tongue body location from 105 to 145 degrees (palatal to uvular) by increments of 10 degrees.

3.2.1.2. Articulator weight modulation. A consonant gesture involves synergistic movements of its associated articulators to achieve its constriction target (see Section 2.2). The weighting parameter for each articulator specifies how much it is engaged in achieving the gesture's target, relative to the other associated articulators. We can introduce some amount of randomness by modulating the articulator weights. Note that the weight parameters are relative values for a given gesture and their relations of any pair to one another should be understood in terms of their ratio. To systematically vary the articulator weights for a given C, we first need to determine the reasonable range of weight ratio sets. The articulators can be divided into *proximal* articulators, which are those most proximal to the effected constriction (lips [UH, LH] for labials, tongue tip [TL, TA] for coronals, tongue body [CL, CA] for velars) and distal articulators (jaw [JA] for labials, jaw [JA] for velars, jaw [JA] and tongue body [CL, CA] for coronals). The weight ratio between proximal and distal articulators was varied in 19 steps where the proximal articulators were the lightest at step 1, the heaviest at step 19, and the mean value of the weight ratios coincided with step 10. The range for the weight modulation was chosen so that an appropriate constriction for a given consonant was ensured, and so that the mean weight ratios (for step 10) were as close to one as possible. Table 6 shows the values of articulator weights used for each consonant. For labials, the mean weight of the upper lip is set slightly heavier than lower lip because the upper lip cannot move down as much as lower lip can move up. For coronals, we shifted the whole range of weight variation to avoid occurrences of non-coronal constriction and hence, the mean ratios of articulator weights are not as close to one as the other consonants. For velars, the mean ratios of CL, CA and JA are all equal at step 10. Since we want to see the effect of constriction location and articulator weight variations separately, we used a default set of constriction locations: 56 degrees and 135 degrees for the constrictions of locations of tongue tip and body, respectively, for the weight modulation.

3.2.2. Results

The generated dataset consists of articulatory trajectories of 300 vowels and either 11 (labial location plus 5 alveolar plus 5 velar) or 57 (19 weights for three locations) consonants from the constriction location and the weight modulations, respectively. All the utterances (bare consonants and vowels) began from a neutral position of articulators and ended when their constriction targets were achieved, which is when the positions of the tongue body constrictions were measured. The Euclidean distance of the constriction positions between C and every V were taken as the index of CV synergy. The smaller the difference they exhibit, the greater the articulator synergy is. The values were normalized so that they range from 0 to 1; then they were subtracted from 1 so that greater synergy would result in a larger number. Each consonant type has synergy values for 300 bare vowels. Thus, a total of 3300 CV synergy values (11 Cs × 300 Vs) from the constriction location modulations and 17,100 values (57 Cs × 300 Vs) from the weight modulation were obtained. We used the same vowel maps (Fig. 4) used in the F/C simulations (Section 3.1) to identify each vowel type. Fig. 6 shows the averaged synergy values for each of the nine places of articulation for consonants (*x*-axis), when combined with the front, central and back vowels (plotted as three functions).

The synergies for the front vowels show a peak at the tongue tip constriction location at 66 degrees (a post-alveolar position), with fairly continuous changes on either side of that. The central vowels have high synergy with the labials, low with the coronals and high again with the velars. There is both gradience and stability within this pattern. The back vowels have high synergy with the labials, consistently low synergy with the coronals, and increasing synergy with the velars as place of articulation becomes more posterior.

The global patterns found for the AP simulation can be seen at all four places of articulation within both the coronal and the velar locations. Although there are some gradual changes within each category, there are clear differences between categories. Further, construction of CV probability ratios using any one of the five coronals and with any one of the five velars exhibit the same pattern: All three diagonals are predicted to be greater than 1.

For the remainder of our discussion, we will average values over the five places of articulation into a single measure for alveolar and velar. Table 7 shows CV ratios normalized from the synergy values for the constriction location and weight modulation, respectively. We used the same formula as in Section 2 to compute a ratio for a specific CV combination, which corresponds to each cell in Table 7

Normalized CV ratio = P(CV)/(P(C) × P(V))
P(CV) = Sum of synergy values for a given CV/Total sum of synergy values
P(C) = Sum of synergy values for a given C/Total sum of synergy values
P(V) = Sum of synergy values for a given V/Total sum of synergy values

The results for all three classifications show that all the diagonals show coronal-front, labial-central, and velar-back preferences are very similar to those in transcribed data (Davis & MacNeilage, 1994, 1995; Davis, MacNeilage, & Matyear, 1999, 2002; Giulivi et al., 2011). Unlike the F/C account, the AP account is also similar to recorded data in ratios for the off-diagonals. For example, the large labial-front ratios obtained with F/C simulation are not obtained here. A detailed analysis with quantitative comparisons to observed ratios will be conducted in Section 4.

The modulation of articulator weight simulations were undertaken to ensure that obtained pattern of results was not specific to a particular assumption about the pattern of articulator cooperation in the production of a given constriction. The synergy values for all from weight values

Table 6
Weight values (in model internal units) at step 1, 10 and 19 for three consonant types: labials (top), velars (mid) and coronals (bottom). Proximal articulators for each consonant are put in bold.

	Step 1	Step 10	Step 19
Labials			
JA	10	1	1
LH	1	1	10
UH	2	2	20
Velars			
JA	10	1	1
CL	1	1	10
CA	1	1	10
Coronals			
JA	500	50	5
CL	500	50	5
CA	500	50	5
TL	1	1	1
TA	1	1	1

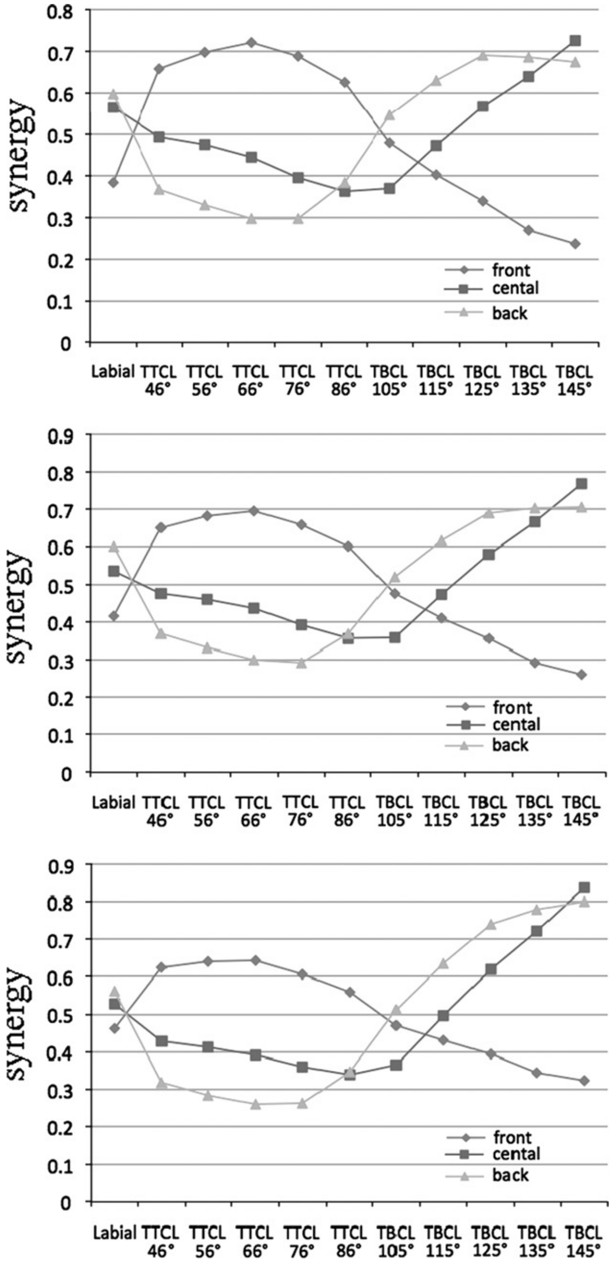

Fig. 6. Synergy values of nine places of articulation for three vowel types (front, central, and back) from the constriction location modulation simulation based on acoustic (top), articulatory (middle), perceptual (bottom) maps.

for a given *C* were averaged and then CV synergy ratios were calculated as they were for the constriction location simulations. The results presented in Table 8 are quite similar to the results of the constriction location simulations, which employed a fixed set of weights. Thus, the obtained patterns of CV synergy ratios appear not to be due to particular weight settings.

4. Comparison of the models

The AP simulation makes the correct prediction that the diagonals values are greater than one, as does the F/C simulation with the adult vocal tract. For the F/C simulation with the infant vocal tract, only two of the three diagonals are greater than 1 (though the third is very close to 1). This suggests that the infant model may be inadequate in some way, which could also account for the poor performance of the F/C account in Serkhane et al. (2007), which also employed an infant vocal tract. Knowledge of the articulatory-acoustic relations may not be yet adequate for this kind of simulation.

Inspection of the off-diagonal values in the babbling data in Tables 1 and 2 reveals some similarity in these values across the two sets of babbling data (MacNeilage & Davis and Giulivi et al.), suggesting that there is some structure in the off-diagonals that has not been described before. Therefore, we decided to compare models based on all nine cells, rather than just the diagonals. The difference between the empirically-obtained ratios and the model values were calculated for all nine cells. For the original empirical data, we averaged the ratios from all published

Table 7
CV synergy ratios from AP simulation of constriction location modulations for acoustic/articulatory/perceptual maps.

	Front	Central	Back
Acoustic			
Coronal	1.27	0.79	0.69
Labial	0.80	1.12	1.19
Velar	0.74	1.19	1.30
Articulatory			
Coronal	1.24	0.74	0.67
Labial	0.86	1.13	1.20
Velar	0.77	1.26	1.31
Perceptual			
Coronal	1.16	0.61	0.59
Labial	0.95	1.11	1.13
Velar	0.83	1.40	1.42

Table 8
CV synergy ratios from AP simulation results from weight modulations for acoustic/articulatory/perceptual maps.

	Front	Central	Back
Acoustic			
Coronal	1.36	0.85	0.82
Labial	0.81	1.18	0.98
Velar	0.62	1.10	1.27
Articulatory			
Coronal	1.41	0.82	0.80
Labial	0.86	1.17	0.94
Velar	0.62	1.06	1.30
Perceptual			
Coronal	1.26	0.91	0.85
Labial	0.67	1.23	1.06
Velar	0.50	1.12	1.37

results from MacNeilage and Davis's babblers (Davis & MacNeilage, 1994, 1995; Davis, MacNeilage, & Matyear, 1999, 2002, see Table 1). Fig. 7 shows the average of the absolute values of the difference scores between the empirical CV probability ratios and either the F/C simulation CV probability ratios (averaged across acoustic, perceptual, articulatory) or the AP CV synergy ratios, separately for the 3 diagonals, the 6 off-diagonals, or all 9. Note that the smaller the difference in the ratios, the closer the results of the simulation are to the empirical data. Three aspects of Fig. 7 are notable: First, the AP synergy account performs noticeably better than the F/C account in estimating the magnitude of the ratios on the diagonal. Second, the off-diagonal accuracy is almost as good. Third, the F/C account, although not having any explicit predictions about the off-diagonals, in fact does better at predicting those than it does with the diagonals in the case of the infant simulations.

To test how well the four models accounted for the *pattern* of reported ratios from babblers across the nine cells (as opposed the absolute error of prediction shown above), we performed a correlation between the ratios obtained by each of the four models separately for acoustic, articulatory and perceptual classifications (Tables 4, 5, 7 and 8), with the ratios reported for MacNeilage and Davis's six babblers (Table 1).

The data scatter plots and resulting regression lines are shown in Fig. 8. The F/C correlations are exceedingly small and none are significant (F/C child: $r=0.06$ ($p=0.88$), $r=-0.37$ ($p=0.33$), $r=0.07$ ($p=0.86$), F/C adult: $r=0.01$ ($p=0.98$), $r=0.05$ ($p=0.90$), $r=0.32$ ($p=0.40$)). AP correlations are modest, and two are significant, while two more are marginal (AP location: $r=0.64$ ($p=0.06$), $r=0.55$ ($p=0.12$), $r=0.43$ ($p=0.25$), AP weight: $r=0.66$ ($p=0.05$), $r=0.66$ ($p=0.05$), $r=0.58$ ($p=0.10$)). The AP models thus matched the pattern across the 9 ratios reasonably well, while the F/C models did not do so at all.

5. Conclusion

Two simulations have shown that there are plausible articulatory sources for the observed preferences in consonant–vowel combinations, both in adult speech and in babbling. Both the Frame/Content (F/C) account and the Articulatory Phonology (AP) were tested with articulatory synthesis. The AP account generated ratios greater than 1 for the preferred combinations that have been found in transcription data. The F/C account did not consistently generate such ratios, and often had velar/back ratios that were much larger than any reported from babbling data. The diagonals only account for approximately half of the observed data, though this cannot be seen from the ratios themselves. Thus the ratios for the off-diagonals should be examined for systematicity. Here, we found that the AP account produced ratios that were much closer to the observed data than the F/C account. Even though the F/C account makes no predictions for the off-diagonals, our F/C simulation resulted in a better match to the off-diagonals than it did to the diagonals for the infant simulation. The error between observed ratios and the AP values, though, were 1/3 the size of the F/C errors.

The F/C models presented here did not generate syllables primarily along the diagonals, somewhat surprisingly as the F/C account predicts most syllables to be on the diagonal. In babbling, approximately 50% of the syllables occur on the diagonals and an equal number off; this was true for both MacNeilage and Davis's and Giulivi et al.'s data (Giulivi et al., 2011). In our F/C child simulation, 46.7% of the syllables were on the

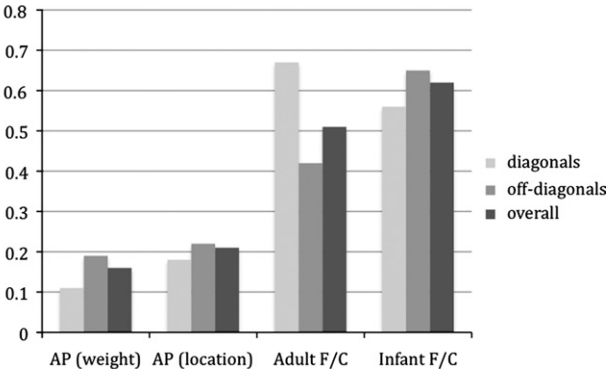

Fig. 7. Differences of ratio magnitude between the simulations and MacNeilage and Davis data.

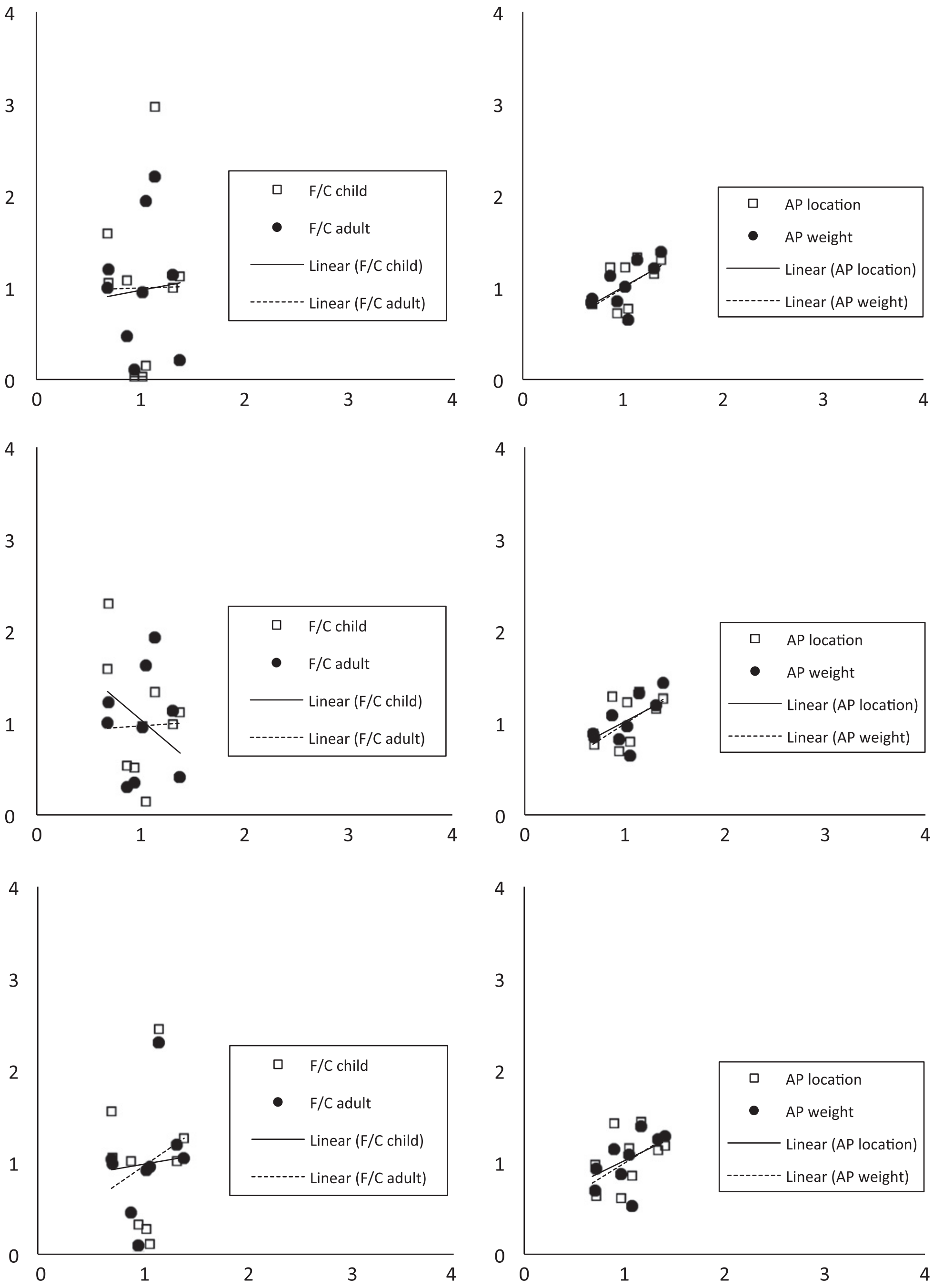

Fig. 8. Regression analysis of FC (left) and AP (right) models against MacNeilage and Davis data based on acoustic (top), articulatory (middle), perceptual (bottom) maps. *X*-axis: Ratios from MacNeilage and Davis data. *Y*-axis: Ratios from models.

diagonal, while 51.3% of the adult F/C simulation were on the diagonal. While this is a relatively good match to the babbling data itself, it contradicts the initial assumption of the F/C model: Jaw motion alone does not result primarily in syllables on the diagonal. Thus there is no evidence in the simulations that jaw motion alone generates almost exclusively the preferred (diagonal) patterns.

The fact that the jaw-only simulations resulted in numerous off-diagonal syllables indicates that the mere presence of such syllables does not imply that infants have direct control over the tongue and lips as articulators. However, the greater match between the AP simulations and the MacNeilage and Davis babbling results provides support for some level of control over more than just the jaw in babbling. The AP model assumes that infants can produce constrictions of lip, tongue tip, and tongue body organs (presumably in an oscillating fashion, though only single constrictions were modeled here), and for each have some control of component articulators that cooperate in achieving those constrictions. Without any such control, as in the F/C simulation, there is no correlation with the data and thus no explanatory value associated with the assumptions behind the account. The AP account holds that infants have control over organ constrictions and that the likelihood of a tongue body position that emerging as a by-products of a given constriction (maximally synergetic vowel gestures) is predictive of that probability of that C–V combination in babbling. Thus a level of control that is greater than just mandibular but less than combinatorial (production of both vowel and consonant gestures) is both plausible and supported by the present results.

The hypothesis that CV preferences have a physiological basis receives support from these simulations. The Articulatory Phonology account matches the observed data more closely than the Frame/Content account. While some caution is necessary in interpreting these findings because the synergies are based on an adult vocal tract (and this should addressed in the future), the fact that the model is so successful in reproducing the off-diagonals gives us some confidence. After all, the fact that it is an adult model did not *a priori* guarantee that it would generate the off-diagonals correctly. The synergy model can also help explain why it is that even in the overall lexicons of adult language, these preferences tend to appear, under the hypothesis that C and V gestures in a CV syllable are triggered synchronously. If the preferences only occurred when control was absent, they should not exist in the adult data. Thus a gesture-based account of speech production can help us understand some of the properties of (seemingly universal) language acquisition.

Acknowledgments

This work was supported by NIH grants DC-00403 and DC-02717 to Haskins Laboratories. We thank Arthur Abramson, Stefan Frisch, Douglas Honorof, Khalil Iskarous, Christine Mooshammer, Aude Noiray, Kenneth J. deJong and four anonymous reviewers for helpful comments.

References

de Boer, B., & Fitch, W. T. (2010). Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, *18*, 36–47.

Boysson-Bardies, B. d., Hallé, P. A., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language*, *16*, 1–17.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Browman, C. P., & Goldstein, L. M. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 151–206.

Browman, C. P., & Goldstein, L. M. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3–4), 155–180.

Browman, C. P., & Goldstein, L. M. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée*, *5*, 25–34.

Davis, B. L. (2010). Speech acquisition. In: W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (pp. 299–315). Malden, MA: Wiley-Blackwell.

Davis, B. L., & MacNeilage, P. F. (1994). Organization of babbling: A case study. *Language and Speech*, *37*, 341–355.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Research*, *38*, 1199–1211.

Davis, B. L., & MacNeilage, P. F. (2004). The frame/content theory of speech evolution: From lip smacks to syllables. *Primatologie*, *6*, 305–328.

Davis, B. L., MacNeilage, P. F., & Matyear, C. L. (1999). Intrasyllabic patterns in babbling and early speech. In: J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences*, Vol. 3 (pp. 2481–2484). San Francisco: University of California, Berkeley.

Davis, B. L., MacNeilage, P. F., & Matyear, C. L. (2002). Acquisition of serial complexity in speech production. A comparison of phonetic and phonological approaches to first word production. *Phonetica*, *59*, 75–107.

Giulivi, S. (2007). *Vowels and consonants favored co-occurrences in language development*. Unpublished Ph.D. Dissertation. University of Florence.

Giulivi, S., Whalen, D. H., Goldstein, L. M., Nam, H., & Levitt, A. G. (2011). An Articulatory Phonology account of preferred consonant–vowel combinations. *Language Learning and Development*, *7*, 202–225.

Goldstein, L. M., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In: M. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 215–249). Cambridge: Cambridge University Press.

Iskarous, K., Fowler, C. A., & Whalen, D. H. (2010a). Locus equations are an acoustic expression of articulator synergy. *Journal of the Acoustical Society of America*, *128*, 2021–2032.

Iskarous, K., Nam, H., & Whalen, D. H. (2010b). Perception of articulatory dynamics from acoustic signatures. *Journal of the Acoustical Society of America*, *127*, 3717–3728.

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

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499–546.

MacNeilage, P. F., & Davis, B. L. (1990a). Acquisition of speech production: Frames, then content. In: M. Jeannerod (Ed.), *Attention and performance XIII* (pp. 453–476). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacNeilage, P. F., & Davis, B. L. (1990b). Acquisition of speech production: The achievement of segmental independence. In: W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 55–68). Dordrecht: Kluwer Academic Publishers.

MacNeilage, P. F., & Davis, B. L. (1993). Motor explanation of babbling and early speech patterns. In: B. d. Boysson-Bardies, S. d. Schonen, P. Jusczyk, P. F. MacNeilage, & J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 341–352). Dordrecht: Kluwer Academic Publishers.

MacNeilage, P. F., & Davis, B. L. (2000). Deriving speech from nonspeech: A view from ontogeny. *Phonetica*, *57*, 284–296.

MacNeilage, P. F., Davis, B. L., Kinney, A., & Matyear, C. L. (2000). The motor core of speech: A comparison of serial organization patterns in infants and languages. *Child Development*, *71*, 153–163.

Maddieson, I. (1984). *Patterns of sounds*. New York: Cambridge University Press.

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

Ménard, L., Davis, B. L., Boë, L.-J., & Roy, J.-P. (2009). Producing American English vowels during vocal tract growth: A perceptual categorization study of synthesized vowels. *Journal of Speech, Language, and Hearing Research*, *52*, 1268–1285.

Ménard, L., Schwartz, J.-L., & Boë, L.-J. (2004). Role of vocal tract morphology in speech development: Perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood. *Journal of Speech, Language, and Hearing Research*, *47*, 1059–1080.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, *53*, 1070–1082.

Mooshammer, C. M., Goldstein, L., Nam, H., McClure, S., Saltzman, E., & Tiede, M. K. (2012). Bridging planning and execution: Temporal planning of syllables. *Journal of Phonetics*, *40*, 374–389.

Nam, H. (2007). Syllable-level intergestural timing model: Split-gesture dynamics focusing on positional asymmetry and moraic structure. In: J. Cole, & J. I. Hualde (Eds.), *Laboratory phonology*, Vol. 9 (pp. 483–506). Berlin, New York: Walter de Gruyter.

Nam, H., Goldstein, L. M., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In: F. Pellegrino, E. Marisco, & I. Chitoran (Eds.), *Approaches to phonological complexity* (pp. 299–328). Berlin: Mouton de Gruyter.

Nam, H., Goldstein, L. M., Saltzman, E., & Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *Journal of the Acoustical Society of America*, *115*2430.

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In: D. Recasens, M.-J. Solé, & J. Romero (Eds.), *Proceedings of the 15th international congress of phonetic sciences* (pp. 2253–2256). Barcelona: Universitat Autonoma de Barcelona.

Oller, D. K. (2000). *The emergence of the speech capacity. Mahwah, NJ*. Lawrence Erlbaum Associates.

Redican, W. K. (1975). Facial expressions in nonhuman primates. In: L. A. Rosenblum (Ed.), *Primate behavior: Developments in field and laboratory research* (pp. 103–194). New York: Academic Press.

Rubin, P. E., Saltzman, E., Goldstein, L. M., McGowan, R. S., Tiede, M. K., & Browman, C. P. (1996). CASY and extensions to the task-dynamic model. *Paper presented at the Proceedings of the 1st ESCA ETRW on speech production modeling and 4th speech production seminar*. Autrans.

Saltzman, E., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*, 333–382.

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

Stoel-Gammon, C., & Herrington, P. B. (1990). Vowel systems of normally developing and phonologically disordered children. *Clinical Linguistics and Phonetics*, *4*, 145–160.

Studdert-Kennedy, M., & Goldstein, L. M. (2003). Launching language: The gestural origin of discrete infinity. In: M. Christiansen, & S. Kirby (Eds.), *Language evolution* (pp. 235–254). Oxford: Oxford University Press.

Tiede, M. K. (1996). An MRI-based study of pharyngeal volume contrasts in Akan and English. *Journal of Phonetics*, *24*, 399–421.

Vihman, M. M., Ferguson, C. A., & Elbert, M. (1986). Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, *7*, 3–40.

Vilain, A., Abry, C., Badin, P., & Brosda, S. (1999). From idiosyncratic pure frames to variegated babbling: Evidence from articulatory modelling. In: J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th international congress of phonetic sciences*, Vol. 2 (pp. 2497–2500). San Francisco: University of California, Berkeley.

von Hapsburg, D., Davis, B. L., & MacNeilage, P. F. (2008). Frame dominance in infants with hearing loss. *Journal of Speech, Language and Hearing Research*, *51*, 306–320.

Vorperian, H. K. (2000). *Anatomic development of the vocal tract structures as visualized by MRI*. Unpublished Ph.D. Dissertation. Madison, WI: University of Wisconsin.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of the Acoustical Society of America*, *117*, 338–350.

Whalen, D. H., Giulivi, S., Goldstein, L. M., Nam, H., & Levitt, A. G. (2011). Response to MacNeilage and Davis and to Oller. *Language Learning and Development*, *7*, 243–249.

Whalen, D. H., Giulivi, S., Nam, H., Levitt, A. G., Hallé, P. A., & Goldstein, L. M. (2012). Biomechanically preferred consonant–vowel combinations occur in adult lexicons but not in spoken language. *Language and Speech*, *55*, 503–515.

Whalen, D. H., Kang, A. M., Magen, H. S., Fulbright, R. K., & Gore, J. C. (1999). Predicting pharynx shape from tongue position during vowel production. *Journal of Speech, Language and Hearing Research*, *42*, 592–603.

Whalen, D. H., Levitt, A. G., & Wang, Q. (1991). Intonational differences between the reduplicative babbling of French- and English-learning infants. *Journal of Child Language*, *18*, 501–516.

Whitney, D. E. (1969). Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man–Machine Systems*, *10*, 47–53.

Zlatic, L., MacNeilage, P. F., Matyear, C. L., & Davis, B. L. (1997). Babbling of twins in a bilingual environment. *Applied Psycholinguistics*, *18*, 453–469.