Development of Articulatory Coordination in Speech Production

音声生成の調音運動の発達

A dissertation presented

by

OOHASHI Hiroki

大橋 浩輝

to

Division of Physical and Health Education

Graduate School of Education

in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the subject of Education

The University of Tokyo

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

General Introduction

Since Eimas *et al.* (1971) [Eimas *et al.* 1971] reported that a month-old infants can discriminate a minimal pair in a manner approximating categorical perception, many language development studies have focused on efficiencies in young children [Jusczyk 1997]. A recent advance in measurement has further enabled us to research neural foundations of language in infancy [Dehaene-Lambertz *et al.*, 2002, Homae *et al.* 2006]. Although studies on speech production have a long history [Jakobson 1968, Lenneberg 1967], the development of speech production remains an open question, even at the level of kinematic articulation. Although this background gives us the impression of children as passive subjects, they must be viewed as autonomous systems whose development is founded on their own spontaneous behaviors, even at the neonatal stage. In this sense, language development is based on autonomous development, which is not necessarily driven solely by external input (Fig. 1.1). One of the aims of this thesis is to elucidate the development of speech production by focusing on articulatory systems. First, I examined the developmental changes of vowel articulation,

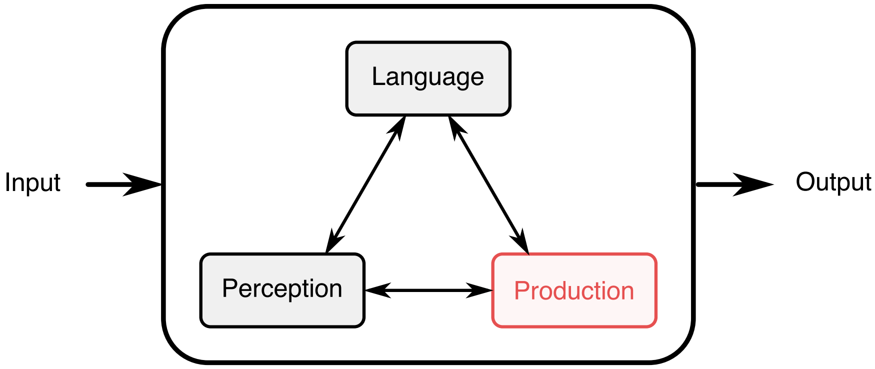

Figure 1.1: Human communication systems and thesis focus.

which comprise relatively stationary movements. Second, I investigated changes in a serial coordination of articulations. I approached these two problems by estimating articulatory movements based on empirically recorded vocalizations by infants.

My approach requires an assessment of anatomy, acoustic analysis, acoustic-articulatory relations, articulatory modeling, and inversion procedures. Therefore, in this chapter I provide an introduction to adult speech production through a discussion of anatomy, motor control, sounds, and perceptions (Secs. 1.1.1, 1.1.2 and 1.1.3). Then, I will review studies on the subject of speech production development, focusing on the articulatory system, the overarching topic of this thesis (Secs. 1.2.1, 1.2.2 and 1.2.3). An articulatory modeling and inversion procedure will be introduced in the methods sections of later chapters (Secs. 2.2 and 3.2).

1.1 SPEECH PRODUCTION

The core matrix of human communication is a verbal interaction. Human beings move their lips, jaws, and tongues to produce vocal sounds following particular rules. When a sound wave propagates through the air, we perceive and decode it to grasp the speaker's intentions. In accordance with this communication process, vocal sounds have traditionally been investigated from three perspectives: acoustic, auditory, and articulatory phonetics. These three perspectives focus on sounds, perceptual cognition, and the system of speech production, respectively. When we produce vowel sounds, we vibrate the glottis (known as *phonation*) and make a certain shape with our vocal tract (known as *articulation*). A change in the stiffness of the glottis results in a change in vibration frequencies, which, in turn, causes a change in pitch perception. Sound waves generated by glottal vibrations pass through the vocal tract, and the resonance caused by the vocal tract emanates as vowel sounds. Based on the source-filter theory [Chiba & Kajiyama 1942, Fant 1960], we can interpret vocal sounds as convolutions of the energy of a sound source, the resonance filter, and the particular property of radiation (Fig. 1.2). As shown in Table 1.1, we convey some types of information by articulation and phonation. At the level of the speech production system, the source of sounds and the resonance filter correspond to the glottal vibration and the shape of the vocal tract, respectively. At the level of sounds, these correspond to a fundamental frequency and formant frequency, respectively; at the level of perceptual cognition, they correspond to prosodic and phonological perception, respectively. In the following sections, I briefly introduce speech production from each of the three perspectives.

1.1.1 Speech production system

First, I will discuss speech production at the level of the speech production system. Anatomically, the human speech production system consists of organs such as the lungs, glottis,

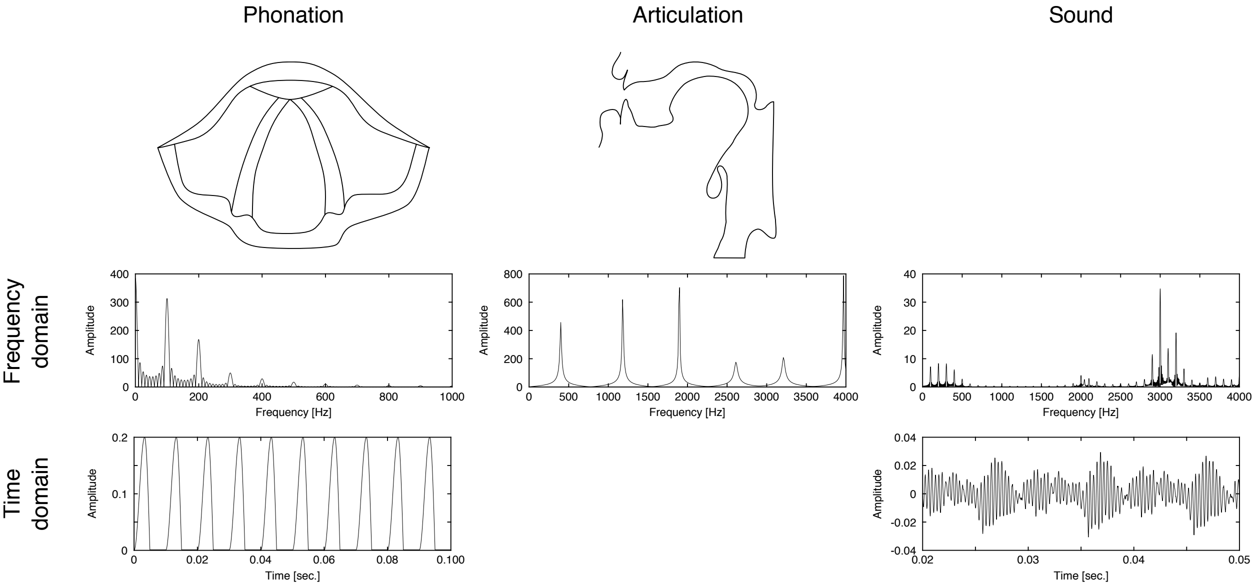

Figure 1.2: According to the source-filter theory [Chiba & Kajiyama 1942, Fant 1960], the sound of a voice can be interpreted as convolutions of energy of sound source, resonance filter, and property of radiation.

Table 1.1: Association of articulation and phonation at levels of related organs, acoustical features, perception, and conveyed information.

	Related Organs	Acoustical features	Perception	Converied information
Articualtion	Tongue, Palate, Jaw, Lips, *etc*	Spectral envelop	Phonological (vowel, consonant)	Linguistic
Phonation	Lung, Larynx	F$_0$, Duration, Power	Prosodic (accent, intonation, pause, *etc*)	Paralinguistic

tongue, lips, and jaw (Fig. 1.3). When the air exhaled from the lungs passes the glottis, the air vibrates the vocal folds at high frequencies; the sounds generated by the vibration are actually the source of the sounds. Fig. 1.4 depicts an idealized cycle of vocal-fold vibration

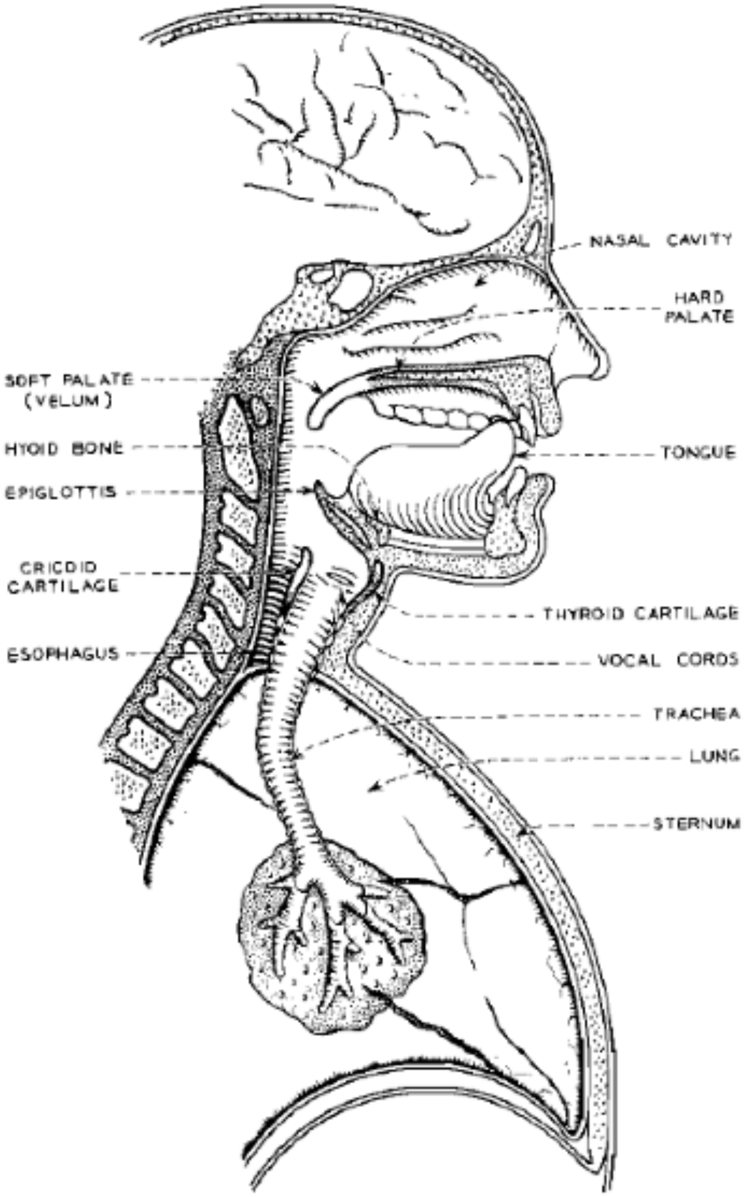

Figure 1.3: Schematic diagram of the human vocal mechanism [Flanagan 1972].

in the coronal plane [Story 2002]. In the first frame of Fig. 1.4, the vocal folds on the left and right sides are initially in contact, and the airway is closed. Next, a lateral movement of the upper portion continues until the left and right sides are separated (Fig. 1.4c). After the air

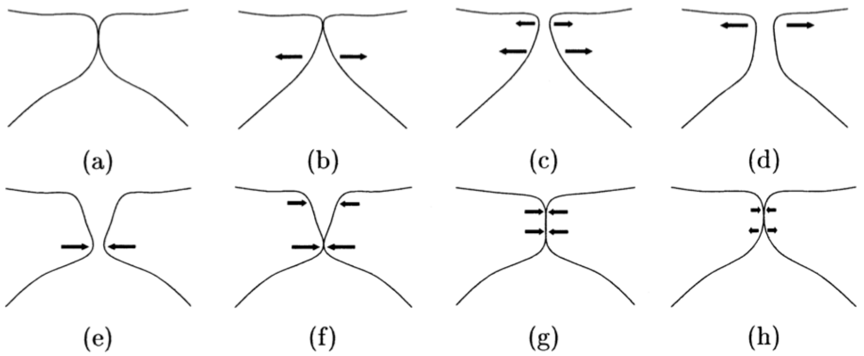

Figure 1.4: Diagram showing an idealized cycle of vocal fold vibration in the coronal plane [Story 2002]. Note that the lower portion of the vocal fold leads the upper portion in creating a wave-like motion on the vocal fold surface, known as a mucosal wave.

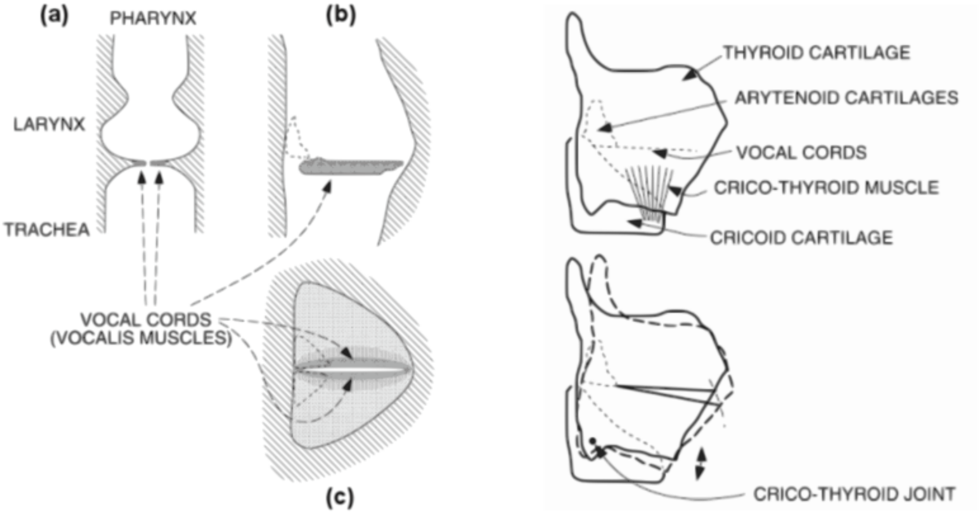

Figure 1.5: Human larynx and two cartilages controlling the vocal folds [Fujisaki 2008]. Left: Sections of the human larynx; (a) anterior-posterior section, (b) median section, and (c) horizontal section. Right: Changes in the relative positions of the thyroid cartilage and the cricoid cartilage due to the activity of the crico-thyroid muscle, causing a change in the vocal cords.

way opens to maximum, the lower portion of the vocal folds begins to move; the upper portion follows (Fig. 1.4e). The entire process repeats cyclically at the fundamental frequency.

Frequency ranges are about 100–200 Hz for male adults, 200–300 Hz for female adults, and 300–800 Hz for young children. The value of the fundamental frequency is subject to the pressure of the air exhaled from the lungs as well as the stiffness of the vocal folds, which is controlled by two cartilages (Fig. 1.5, right panel), the thyroid and cricoid cartilages, whose relative positions are altered by crico-thyroid muscle activity. This phonatory system enables us to generate paralinguistic information, defined as information that is not inferable from written characters but is intentionally added by the speaker to express linguistic information (*e.g.*, accent and intonation).

Figure 1.6 shows the human articulatory system. This system consists of numerous organs, including the jaw, lips, teeth, tongue, nasal cavity, palate, velum, and pharynx. Among these, the tongue plays a major role in producing vowels and consonants. Although a multitude of muscles govern the articulatory movements of as many as 100 muscles [Kent 2004], muscles involved in the tongue movement can be divided into two categories: extrinsic and intrinsic muscles. In addition, the tongue is traditionally divided into three portions: the apex, the dorsum, and the body (Fig. 1.7). Extrinsic muscles (including the styloglossus, genioglossus, hyoglossus, and others) are located on surface of the tongue and attach the tongue to other structures (Fig. 1.7). Extrinsic muscles largely influence the shape of the tongue, with the three aforementioned extrinsic muscles responsible for tongue movements employed in vowel production. Articulatory movements controlled by the position of the tongue body are achieved by the posterior genioglossus and hyoglossus, while movements controlled by the shape of the tongue dorsum are achieved by the anterior genioglossus and middle styloglossus [Honda 1996, Takano & Honda 2007]. On the other hand, the intrinsic muscles, which include longitudinal muscles, verticalis, and transverse muscles, are entirely

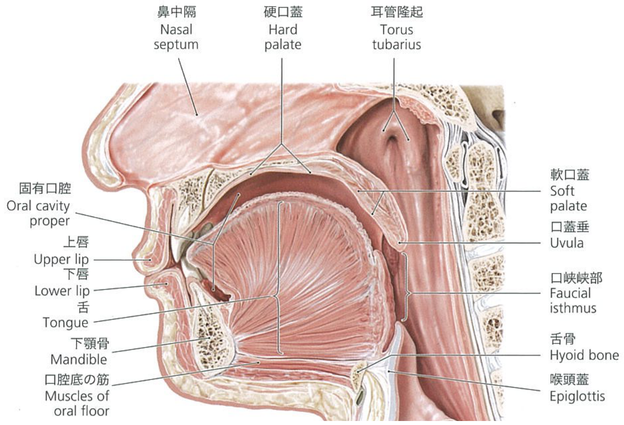

Figure 1.6: Midsagittal view of the anatomical structure of the vocal tract [Schünke *et al.* 2007].

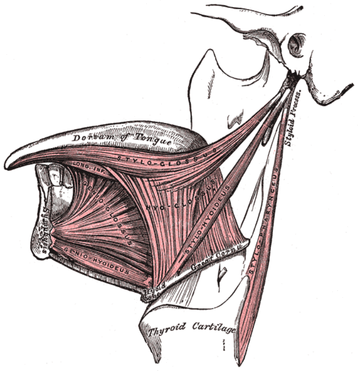

Figure 1.7: Structure of human tongue [http://www.bartleby.com/107/]

located within the tongue and alter its shape. For instance, during production of the /r/ consonant, the superior longitudinal muscle flicks up the tongue apex.

Here, I associate articulation with vowels and consonants. In brief, articulatory movements are close-open alternations of the vocal tract. The opening and closing of the vocal tract is associated with vowels and consonants, respectively. Vowels are mainly characterized by the horizontal and vertical positions of the tongue (Fig. 1.8): according to the horizontal positions of the tongue, vowels are classified into front, center, and back vowels; according to the vertical positions, they are classified into high, middle, and lower vowels. Consonants can be classified into categories according to their places of articulation (Table 1.2). Of these, the labial, coronal, and dorsal consonants are major categories. Consonants are also classified by the manner of articulation, such as the fricative, nasal, and stop. Consonants are produced as a result of the coordinated movements of one or more articulators; these movements are referred to as consonantal gestures. The lips, tongue apex, and tongue dorsum articulate to produce labial, coronal, and dorsal consonants, with the jaw potentially contributing to all of these consonants. Both coronal and dorsal consonants are produced by the tongue but are governed by different tongue tissues. These redundancies allow for flexible coordination and the generation of stable speech production even if sudden perturbation prevents the movements of any one organ [Kelso *et al.* 1984, Trembley *et al.* 2003].

Figure 1.8: IPA vowel chart. Black color chart and red circles indicate IPA vowel chart and Japanese vowels, respectively.

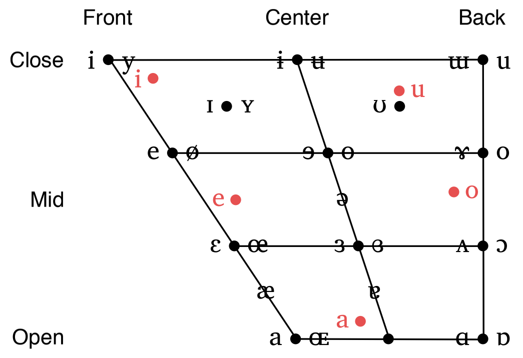

Table 1.2: IPA consonant chart.

	Bilabial	Lab. dent.	Dental	Alveolar	P-alveo.	Retroflex	Palatal	Velar	Uvular	Pharyng.	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ	▨	ʔ
Nasal	m	ɱ		n		n	ɲ	ŋ	ɴ	▨	▨
Trill	ʙ			r				▨	ɾ		▨
Tap/Flap				ɾ		ɽ		▨			▨
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lat. fric.	▨	▨		ɬ ɮ						▨	▨
Approx		ʋ		ɹ		ɾ	j	ɰ			▨
Lat. appr.	▨	▨		l		ɭ	ʎ	ʟ		▨	▨

To produce a sequence of phonemes such as syllables, we must exhale air, vibrate the vocal folds, and move many articulatory organs in appropriate timing. Relatively simple movements, such as close-open alternation of the vocal tract mainly achieved by the rhythmic mandibular movements, are thought to be controlled by the central pattern generators located in the brain stem. More complex articulatory movements that require coordination among many organs are founded on cerebral cortex. These flexibly coordinated behaviors are controlled by somatotopically arranged yet partially overlapping articulator representation on ventral pre- and post-central gyri, which activate in a spatio-temporally organized manner [Bouchard *et al.* 2013]. Language production involves phonological, lexical, and grammatical information; Broca's area conducts a sequencing process to organize them [Sahin *et al.* 2009].

Many phonological studies have explored the processing of linguistic information. Based on traditional phonological perspectives, phonotactics is also governed by a set of universal rules. Phonologists take two main approaches — Principles-and-Parameters approach to Universal Grammer (P&P) [Jackendoff 2002] or Optimality theory (O.T.) [Prince & Smolensky 2004] — to seek the rules. According to P&P, linguistic structures rely on a set of universal rules and parameters; switching the parameters leads to changes in linguistic structures. On the other hand, O.T. states that the ranking of rules decides linguistic structures. For instance, variations between two pronunciations of the English plural suffix can be explained by the rule devoicing a suffix after voiceless consonants. Based on O.T., a plural sequence (output) is generated from stem-*s* (input) as follows: two sibilants cannot be adjacent, input segments in output segments should be maximized, voice specifications of the left edge of every plural morpheme must coincide with the specification of the right edge of a noun stem, output segments should be contained in input segments, and voice specification should be maintained. One indicator of the existence of a set of universal rules is the parallelism between natural sign language and spoken language [Jackendoff 2002]. Since their systems differ completely, neural representations of sign and spoken language should be different from each other. However, a classical lesion study [Hickok *et al.* 2001] reported that left-hemisphere language dominance may be observed in both spoken language and natural sign language, and this report have been confirmed by a following study [Petitto *et al.* 2000]. These results suggest that language is independent of its modality.

1.1.2 Sounds of speech

In this section, I will discuss speech production at the level of sounds. As I described in Sec. 1.1, we can divide sound characteristics into two components: the sources of sounds and the resonance filters. At the level of sounds, each component corresponds to a fundamental and formant frequency. From the standpoint of physical acoustics, using the filter's frequency response we can obtain a coarse geometrical shape of the vocal tract (assuming that sound is a stationary wave, nodes and anti-nodes determine a wave's resonance frequency). In conclusion, the degree of vertical (close-open) and horizontal displacement (front-back) corresponds to the first and second lowest resonance frequencies, respectively (hereafter, first formant frequency F_1 and second formant frequency F_2). Based on these links, we can depict a graphical representation of vowels in terms of F_1 and F_2 (Fig. 1.9). Note that we cannot always associate changes in geometrical shape of the vocal tract with movements of a single articulatory organ. For instance, geometrical changes caused by movements of the jaw par-

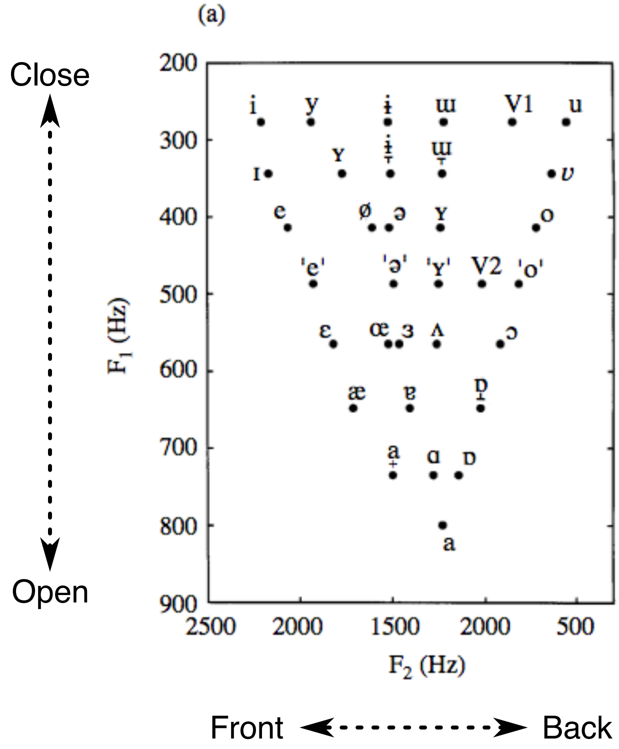

Figure 1.9: Acoustic prototypes in the (F_1, F_2) space (in Hz) [Schwartz *et al.* 1997].

tially overlap with those caused by movements of the tongue body. Furthermore, a signal processing procedure called linear predictive coding (LPC) [Atal & Hanauer 1971, Itakura 1975] enables us to extract the filter's frequency response from empirically recorded sounds. On the other hand, a fundamental frequency can be estimated by using autocorrelations, instantaneous frequencies, and other methods.

1.1.3 Perceptual cognition

Finally, I discuss speech production by way of perceptual cognition. Although there exist many studies dealing with auditory perception, I discuss only those involved in speech production. It is well known that auditory perception affects speech production. For instance, previous studies [Yates 1963] have shown that delayed auditory feedback perturbs our speech production. Moreover, distortion of formant frequencies in auditory feedback may influence the formant frequencies of online speech production [Houde & Jordan 1998]. Note that perception does not have a unilateral effect on speech production—it is a bilateral interaction. Speech production's influence on perception has been investigated from the perspective of the speech motor theory [Liberman 1957, Liberman *et al.* 1967, Liberman & Mittingly 1985, Liberman & Whalen 2000]. In brief, this theory claims that (i) speech perception is special because it is achieved through a different process than auditory processing, (ii) speech perception entails the perception of a gesture, and (iii) speech perception involves a motor process. At the behavior level, studies investigating the McGurk effect [McGurk & MacDonald 1976], duplex perception [Mann & Liberman 1983], categorical perception of phonemes [Liberman 1957] and verbal transformation effect [Sato *et al.* 2006] have been traditionally adduced as evidence for the speech motor theory. Although its claims are controversial [Galantucci *et al.* 2006], recent neuroscientific results support a core theory claim (iii). For instance, by using transcranial magnetic stimulation, previous studies have reported that, during speech perception of words that heavily involve tongue movements, there is an increase in motor-evoked potentials at the tongue muscles [Fadiga *et al.* 2002]; additionally, it has been demonstrated that repetitive transcranial magnetic stimulation applied over the ventral premotor cortex delays phoneme discrimination requiring phonemic segmentations [Sato *et al.* 2009]. Furthermore, the verbal transformation effect is not only founded on activities of the superior temporal cortex but also those of the anterior cingulate and inferior frontal cortices, suggesting that neural representations of the motor system are recruited into auditory perception [Kashino & Kondo 2012, Kondo & Kashino 2007]. Above all, this theory reminds us that cognition is not an isolated process, but is embedded into ecological environments and percepto-motor interactions.

1.2 DEVELOPMENT OF SPEECH PRODUCTION

The development of speech production during the first years of life has been characterized as following a specific course. Infants are born able to produce spontaneous sounds, such as sneezing and crying. They then produce cooing (*i.e.*, quasivocalic sounds similar to vowels). Subsequently, coos expand into clear vowel sounds characterized by their full resonance and wide variety. At an early stage of babbling, infants repeat the same consonant-vowel units such as "papapa" and "mamama." Finally, beginning around the end of the first year of life, infants produce meaningful speech. In Sec. 1.1, I introduced adult speech production. Next, I will review how speech production systems change throughout development, mainly focusing on young children.

1.2.1 Developmental changes in speech production system

Previous studies [Fitch & Giedd 1999, Goldstein 1980, Sasaki *et al.* 1977, Vorperian *et al.* 1999, Vorperian *et al.* 2005] on the anatomy of the infant vocal tract reveal that the shape of children's vocal tracts varies significantly from those of adults. For instance, Voperian *et al.* (2005) [Vorperian *et al.* 2005] examined the growth pattern of the vocal tract structures, visualized using magnetic resonance imaging, and assessed their longitudinal changes in terms of tract length throughout the first six years of life. As shown in Figs. 1.10A and C, the length of the vocal tract at 6 months of age is less than 8 cm, about half that of an average adult. I nfants' vocal tracts are not only smaller than those of adults, but they have a broader oral than pharyngeal cavity, and a more gradually sloping pharyngeal tract. The broader oral cavity is shown in Figs. 1.10A, D and E. In general, a broader oral cavity creates a larger ratio of vocal tract horizontal length to vertical height. Figs. 1.10D and E demonstrate the developmental changes of vertical height and horizontal length of the vocal tract, respectively. While this ratio is about 9/10 [cm/cm] in male adults, that of newborns is about 5/3 [cm/cm], with pharyngeal cavity size becoming larger due to laryngeal descent throughout development. Other studies [Lieberman 1969, Lieberman 2012] focused on this descent and its concomitant reshaping of the tongue. Researchers claimed that newborns' tongues are flat and located almost entirely in the oral cavity, akin to chimpanzees, and that laryngeal descent enables us to produce the *quantal* vowels /i/, /u/, and /a/. Descent is thought to occur in two steps: first, the laryngeal skeleton descends in relation to the hyoid bone; second, the hyoid bone descends in relation to the mandible. The second descent does not occur in chimpanzees [Nishimura *et al.* 2003], and the hyoid bone's proximity to the base of the tongue prevents human-like speech production.

Laryngeal descent would also influence the arrangement of the tongue tissue. A previous study [Takemoto 2008] compared human tongue muscles to those of chimpanzees. The inner muscles of a chimpanzee's tongue were revealed to be oriented in parallel to one another, as opposed to those of an adult human. Because the direction of force is the same throughout most of the tongue, most of its parts will have a lesser degree of freedom of deformation. Human infants' tongues may also have a lesser degree of freedom for a similar reason.

A prevailing view is that the laryngeal descent allows human beings to be more flexible

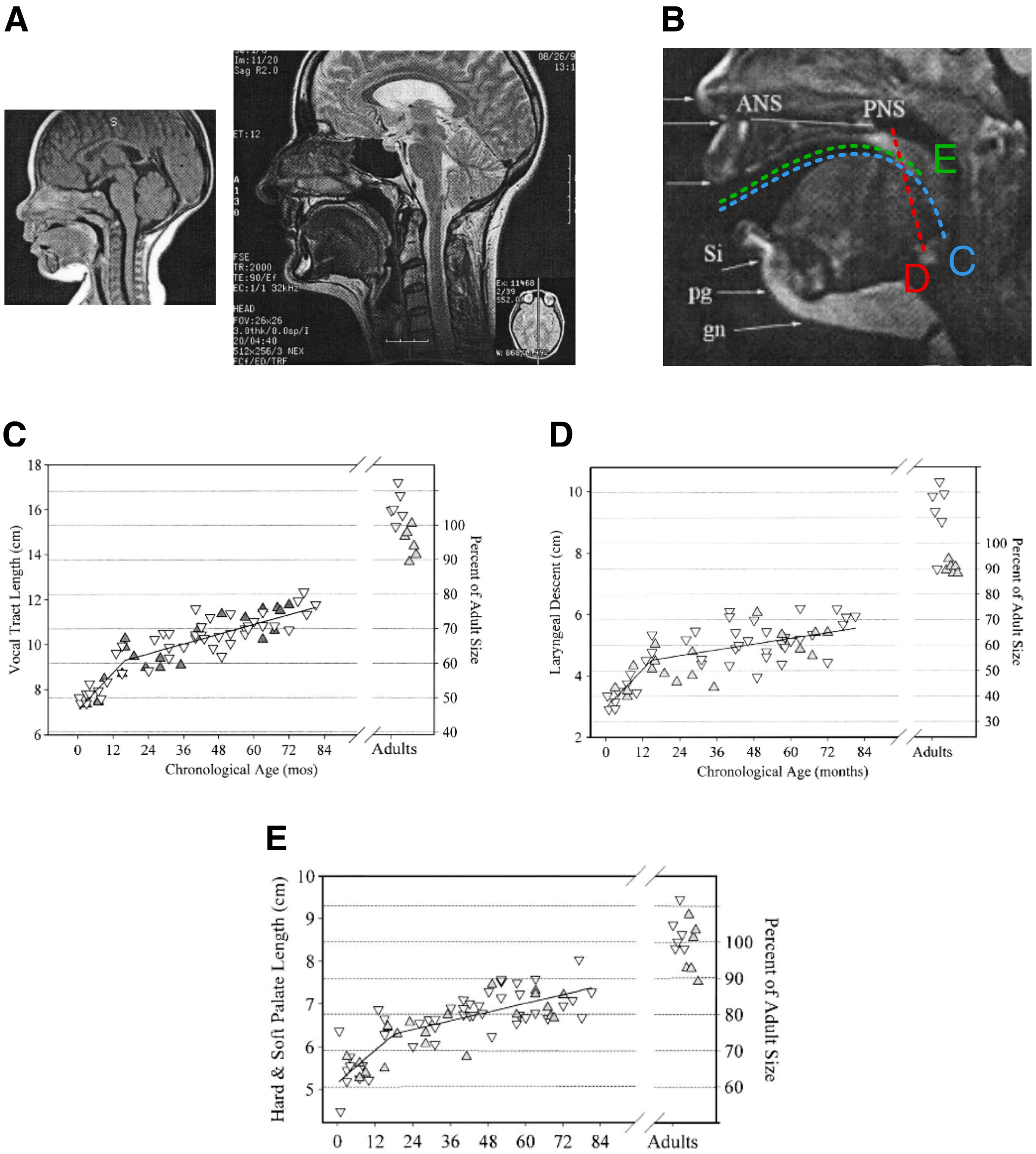

Figure 1.10: Development of the vocal tract structure [Vorperian *et al.* 2005]. A: Midsagittal magnetic resonance images of a 7-month-old female (left) and adult female (right). B: Each definition of lengths in C–E. This figure was generated by clipping and modifying the original figure. Vocal tract length (C), laryngeal descent (D), and hard and soft palate length (E) of pediatric and adult cases (open triangle down for males, and shaded triangle up for female). The second Y-axis reflects the percentage of adult size.

in speech production. However, a recent advance in measurement technique raises questions about the role of laryngeal descent. Recent studies [Boë *et al.* 2007, Boë *et al.* 2013] have insisted that, despite the great difference between infant and adult vocal tracts, limitations in infancy are a matter of control rather than anatomy. In a series of studies, these researchers compared empirically measured acoustical features with those simulated by the model and found them to be compatible. Moreover, they confirmed that tongue fibers in newborns are quite similar to those of adults (Fig. 1.11).

Next, I will review the development of the phonatory system. As shown in Fig. 1.12A [Monnier 2011], from birth to adolescence, the diameter, length, and cross-sectional area of the trachea increase two-, three-, and sixfold, respectively. Infants' larynxes differs from those of adults as follows (Fig. 1.12B) [Monnier 2011]: (i) the epiglottis is omega-shaped; (ii) the lateral edges of the epiglottis are positioned slightly medial to the pharyngo-epiglottic fold; (iii) the aryepiglottic folds are shorter; (iv) the increased ratio of the cartilaginous to ligamentous glottis accentuates the pentagonal shape of the glottis during inspiration; (iv) the immediate subglottis lumen is elliptical, due to the V-shaped upper half of the cricoid cartilage.

Although many studies have examined the anatomical development of the speech production system, relatively little research has focused on the development of motor control in the speech production system, especially during the first year of life. Articulatory movements of adults can be measured by radiographic imaging, electromagnetic articulography and electropalatography, magnetic resonance imaging, ultrasound, and motion-capture systems. Phonatory movements, or the vibration of the vocal folds, can be captured by electroglot-

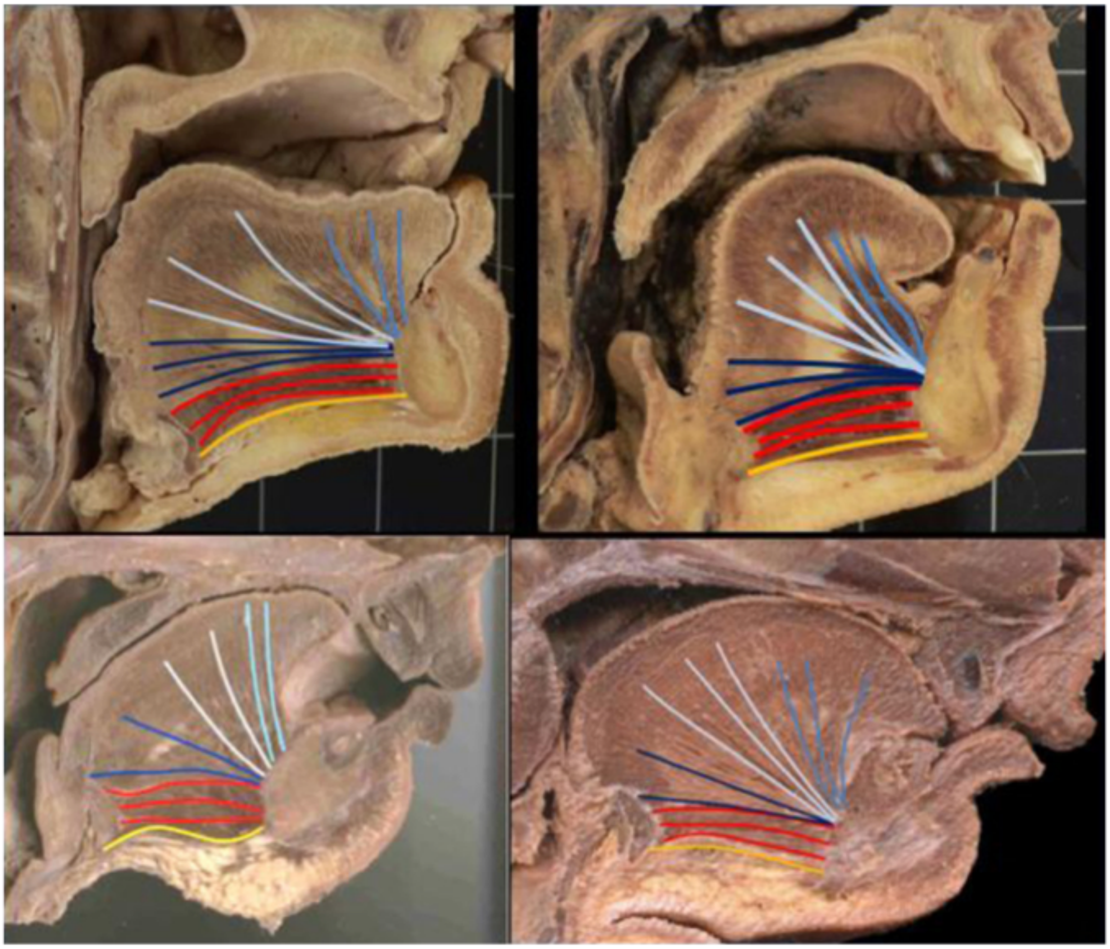

Figure 1.11: The intrinsic muscle fibers of the tongue [Boë *et al.* 2013]. On the bottom row, two fetuses shortly preterm (31–32 weeks amenorrhea on the left, 35–36 weeks amenorrhea on the right), and on the top row two adults.

tography and fiberscope. However, because of ethical issues and the difficulty in prompting specific vocalizations, typical methods of measuring articulatory or phonatory movements cannot be applied to young children. Nonetheless, it is possible to measure external effectors such as the jaw and the lips. Previous studies [Green *et al.* 2000, Green *et al.* 2002, Nip *et al.* 2009] measured displacement of the jaw and lips during speech produced by infants over 9 months of age. They reported that the development of coordination between the jaw

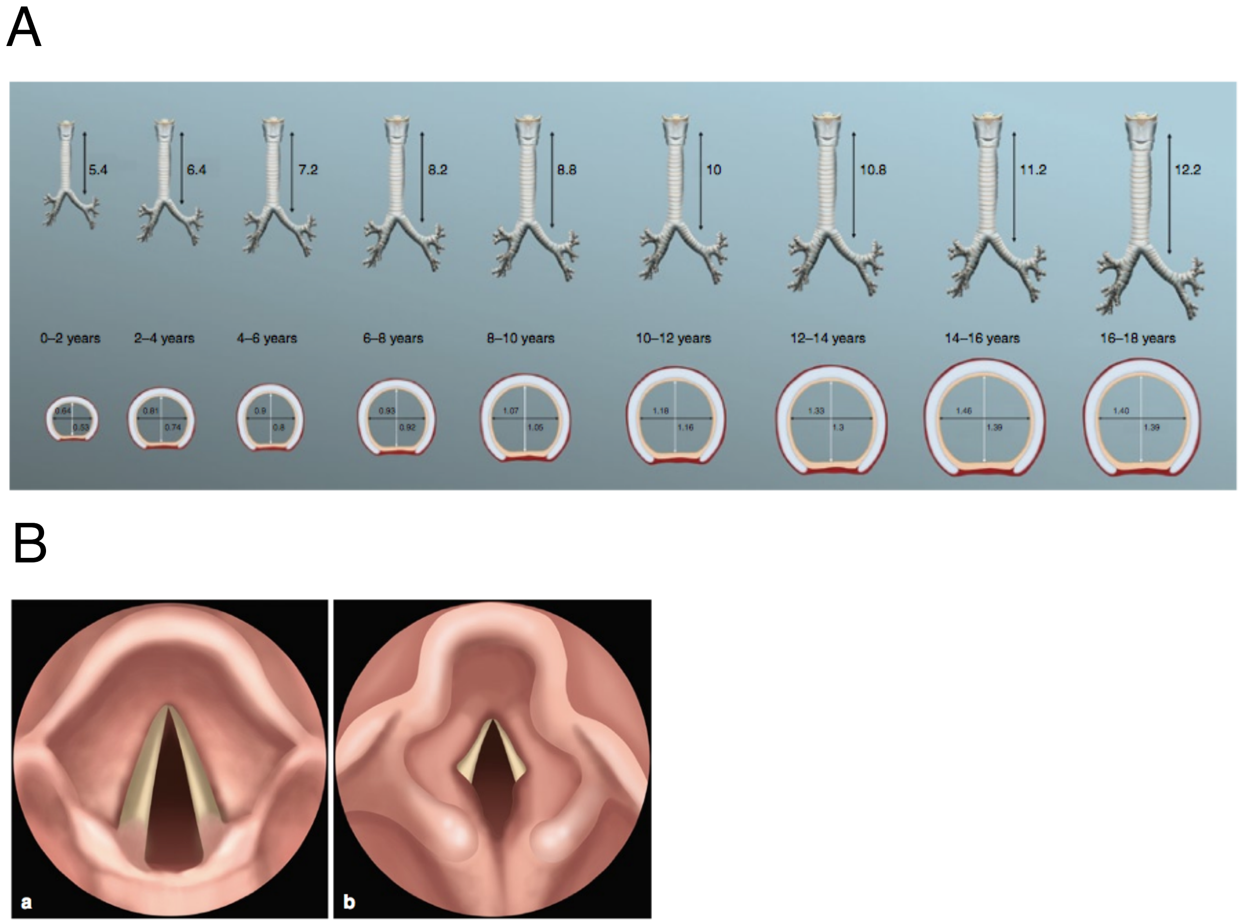

Figure 1.12: (A) Tracheal lengths and diameters: from birth to adolescence, the length of the trachea doubles, its diameter triples, and its cross-sectional area increases sixfold. (B) Schematic endoscopic aspect of the adult and infant larynx [Monnier 2011].

and lips necessary for speech goes through three primary phases: integration, differentiation, and refinement [Green *et al.* 2000, Green *et al.* 2002]. They also found that jaw movements mature earlier than lip movements [Nip *et al.* 2009].

1.2.2 Developmental changes in sounds of speech

As I described in the beginning of Chap. 1 numerous studies have assessed the development of speech production. Employing transcriptions and acoustic features, these studies have revealed characteristics of early vocalizations and inferred the underlying articulatory mechanisms.

Classically, three claims have been made about speech production development [de Boysson-Bardies 1999, Jakobson 1968, Oller 2000]: (i) infants produce all possible sounds of all languages with ease, (ii) there is no phonetic relationship between babbling and speech, and (iii) infants produce babbling sounds at random. However, many studies have shown that these neither characterize the first stage of infant vocalization, nor do they accurately describe canonical babbling [Locke 1983, MacNeialge 2008, Oller 2000]. The inaccuracy of these claims derives in part from transcription analysis. For instance, infants at one month of age rarely produce linguistic segments such as phonemes or syllables, and sounds at this stage are produced with the vocal tract in a relaxed breathing posture. Therefore, it makes no sense to infer well-formedness based on transcription analysis in infancy.

Oller (2000) [Oller 2000] proposed the infraphonological approach, taking into account the bias of well-formedness caused by transcription analysis. Using this approach, he attempted to determine the extent to which infant sounds revealed command of the principle of well-formed speech sound construction. Oller (2000) [Oller 2000] defined early speech development production as shown in Table 1.3. At the phonation stage, infants produce not only brief vocalizations, called quasivowels, but also sneezing, crying and laughter sounds. Quasivowels lack a certain articulatory posture, and are produced with the vocal tract at

Table 1.3: Stages of infant vocal development as seen through infraphonological interpretation [Oller 2000]

Names of stages	Onset ages in months	Protophones mastered	Infraphonological principles mastered
Phonation	0–2	Quasivowels	Normal phonation
Primitive Articulation	1–4	Gooing	Articulation
Expansion	3–8	Marginal babbling, full vowels, respberries, squealing	Full resonance
Canonical	5–10	Canonical babbling	Rapid formant transition

rest. Infants at this stage may explore a variety of phonations to find all modes of vocal fold oscillation. This exploration underlies paralinguistic communications [Buder *et al.* 2008]. At the primitive articulation stage, the dorsum of the back of the tongue comes into contact with the back of the throat or palate, and generates cooing. At the expansion stage, infants actively move articulatory organs to explore possibilities of the vocal tract posture and associate these postures with the sounds they generate. At the canonical stage, infants produce closure and opening sequences with normal phonation. In particular, infants at this stage produce repetitive sequences such as /papa/ and /mama/.

Contrary to the salience of a repetitive sequence at the canonical stage, repetitive sequences play only a minor part in many languages. Throughout development, children also acquire capacities to produce variegative sequences. MacNeilage and colleagues [MacNeilage & Davis 2000, MacNeialge 2008] have revealed a shared preference for specific articulatory patterns (Fig. 1.13). They focused on three kinds of consonants preferred from early development: labial, coronal, and dorsal. The group analyzed the phonotactics of babbling and first words, and reported two findings: first, although there are nine possible CV sequences from a theoretical standpoint, infants at the babbling stage prefer three specific CV patterns (labial-center, coronal-front, and dorsal-back combinations) at significantly higher frequencies (Fig. 1.13A); second, at the beginning of the transition from repetition to variegative productions, labial-vowel-coronal CVC sequences are preferred to coronal-vowel-labial ones (Fig. 1.13B). Researchers claimed that these preferences reflect an important role of the jaw in the serial order production of early development. In their theory, known as the Frame/Content theory, they claimed that three *Frame* underlying mature syllable structures can be observed in CV sequences. Among three frames, they especially emphasized the labial-central CV sequences, which are referred to as the *pure frame*, and proposed that a pure frame is generated by rhythmic oscillations of the jaw with the tongue at a rest. Throughout development, active tongue movement is recruited into the oscillations to fill the *Content* of the pure frame.

Acoustic features of children's speech sounds have also been extensively studied. As I described in Sec. 1.1.2, vocal tract shape can be inferred from formant frequencies. Based on these associations, many previous studies [Ishizuka *et al.* 2007, Kent & Murray 1982, Vorperian & Kent 2007] have examined the longitudinal changes in child vocalization formant frequencies and reported that formant frequencies become lower and more distinct in vowel type as children mature.

For instance, Ishizuka *et al.* (2007) [Ishizuka *et al.* 2007] examined longitudinal changes

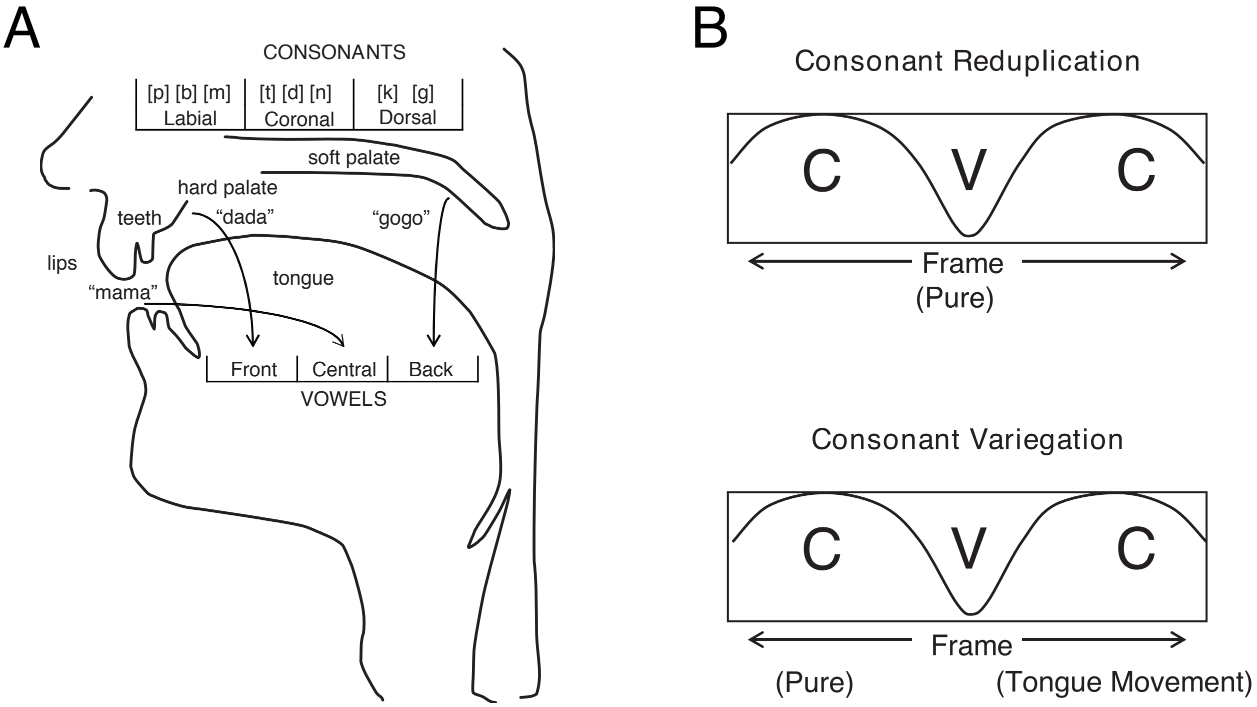

Figure 1.13: The Frame/Content theory proposed by MacNeilage and colleagues [MacNeialge 2008]. (A) Schematic view of the articulatory component of speech showing the three favored consonant-vowel (CV) co-occurrence patterns. (B) The conceptualizations of the development of the labial-coronal sequence effect.

in the formant frequencies of vowel sounds included in the NTT infant database, which I also used in the analysis in Chaps. 2 and 3. Although cross-sectional or longitudinal approaches are possible for data collections, the number of subjects, samples, or the range of months is important in both methods. To overcome these problems, in the CHILDES project [MacWhinney 2000], scholars shared data to develop an infant speech database. However, the Japanese database employed in the project may not be suitable for a longitudinal analysis because collection periods are short, the sample size is not large and utterance files are not provided. To remedy this limitation, Amano *et al.* (2009) [Amano *et al.* 2009] constructed the NTT infant database, digitally recording utterances of five infants and their parents, and attaching detailed information such as session, utterance, transcription, property tag, time record, comment, fundamental frequency, voice/unvoiced label, and phoneme label files. I will only address the utterance, property tag, comment, fundamental frequency, and phoneme label files, which I use in Chaps. 2 and 3. The utterance files are generated to extract each utterance, segmented by a silent period of more than 500ms. Utterances are digitized with 16-bit quantization at a 16-kHz sampling frequency and saved in the WAV format. During recordings, children and their parents were not required to complete particular tasks apart from recording their talk during everyday situations. The property tag files are generated to identify speakers (infant, father, mother, or other person), listeners (infant-directed or adult-directed) the background noise level (low or high), and the utterance loudness level (low, normal, or high). The comment files are generated to identify characteristics of utterances (laughing, hiccupping, coughing, sneezing, yawning, crying, singing, reading aloud, number counting, babbling, *etc*). The fundamental frequency files contain values of fundamental frequencies estimated using the Ripple Enhanced Power Spectrum method [Nakatani & Irino 2004]. The phoneme label files include phonemic transcriptions of utterances and the start and end times of each phonemic segment, measured in milliseconds. By analyzing the NTT infant database, Ishizuka *et al.* (2007) showed that distances between centroids of acoustical distribution of vowels become larger as children mature, suggesting that clusters of vowels become more distinct throughout development (Fig. 1.14). The lower formant frequencies may be interpreted as developmental changes in the length of the vocal tract, because it elongates as formant frequencies become lower. The trend that formant frequencies of each vowel become more clearly different from the other vowels implies a differentiation process of articulatory states during vowel production.

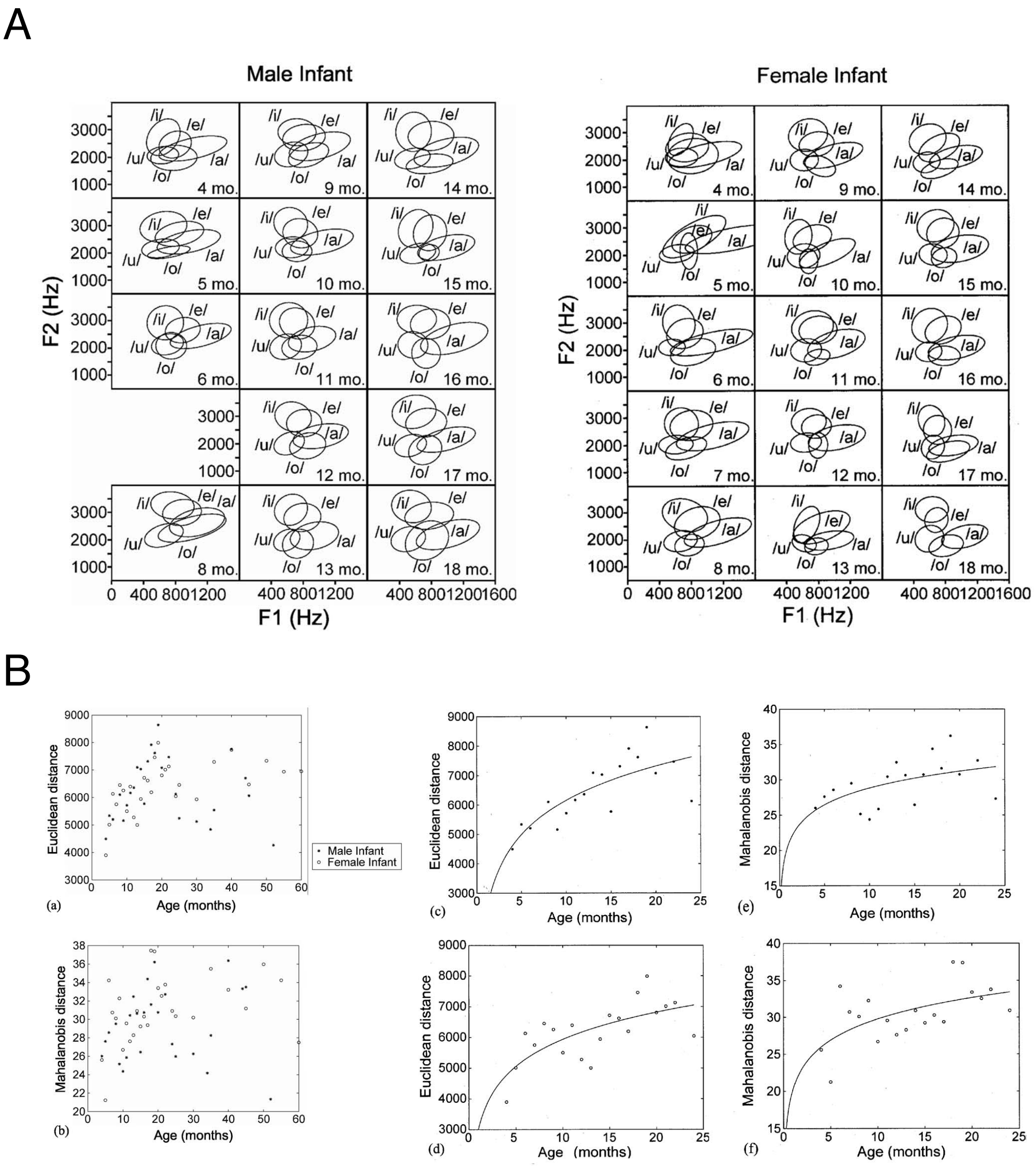

Figure 1.14: (A) Fifty percent probability ellipses for each vowel on the F_1 and F_2 plane by month of age. (B) Left column: the sum of the Euclidean distances between the vowel centroids for each month of age. Middle column: the sum of the Mahalanobis distances for male (top) and female (bottom) infants between vowel distributions for each month of age. Right column: the sum of the Euclidean distances for male (top) and female (bottom) infant between vowel distributions for each month of age [Ishizuka *et al.* 2007].

Contrary to vowels, consonants are characterized by dynamic features such as locus of transition of formant frequencies and voice onset time (VOT). Studies exploring locus [Gibson *et al.* 2007,Sussman *et al.* 1996] , which reflects a degree of CV sequence coarticulation, have reported that development of coarticulation patterns varies from consonant to consonant. Other researchers [Whalen *et al.* 2007] focused on VOT, which signals voicing contrast (*e.g.*, between /p/ and /b/) in the babbling of English and French infants at 9 and 12 months. They reported the possibility that infants can control prevoicing more successfully than aspiration.

Numerous studies have addressed longitudinal changes in acoustic characteristics of vocalizations produced by children, contributing to our understanding of the development of articulation. However, because we cannot consistently associate changes in the geometrical shape of the vocal tract with movements of a single articulatory organ, a detailed explanation of the development of articulatory movements has yet to be formulated.

1.2.3 Developmental changes in perceptual cognition

Development of perceptual cognition has been extensively examined by many previous studies [Jusczyk 1997,Kuhl 2004] (Fig. 1.15). Typical topics include perception of minimal pairs and word segmentation. In general, universal phonetic perceptions localize into perceptions suitable to a specific language, and prominent accent patterns and statistical learning facilitate infants' segmentation of linguistic units embedded in a continuous speech (Fig. 1.15).

Besides topics of purely auditory perception, one instance of perceptual cognition involvement in speech production is imitation. Contrary to classical cognitive development theory [Piaget 1969], which claims that imitation is only observed in children older than two years, Meltzoff and Moore (1977) [Meltzoff & Moore 1977] reported that infants under 1 month of age imitated oral gestures and directed questions. Other research [Kuhl & Meltzoff 1982] reported that neonates recognize the correspondence between auditorily and visually presented speech information. Moreover, some infants have been observed to imitate vowel sounds presented during experiments [Kuhl & Meltzoff 1996]. Although the existence of neonatal imitation is a controversial question [Jones 2007], some believe that neonates can cross-modally process visual, auditory, and motor information, and represent the equivalent motor response [Meltzoff & Moore 1977]. In this sense, neonatal imitation may be one of entrainment among the visual, auditory, and motor domains, and vary from the imitation

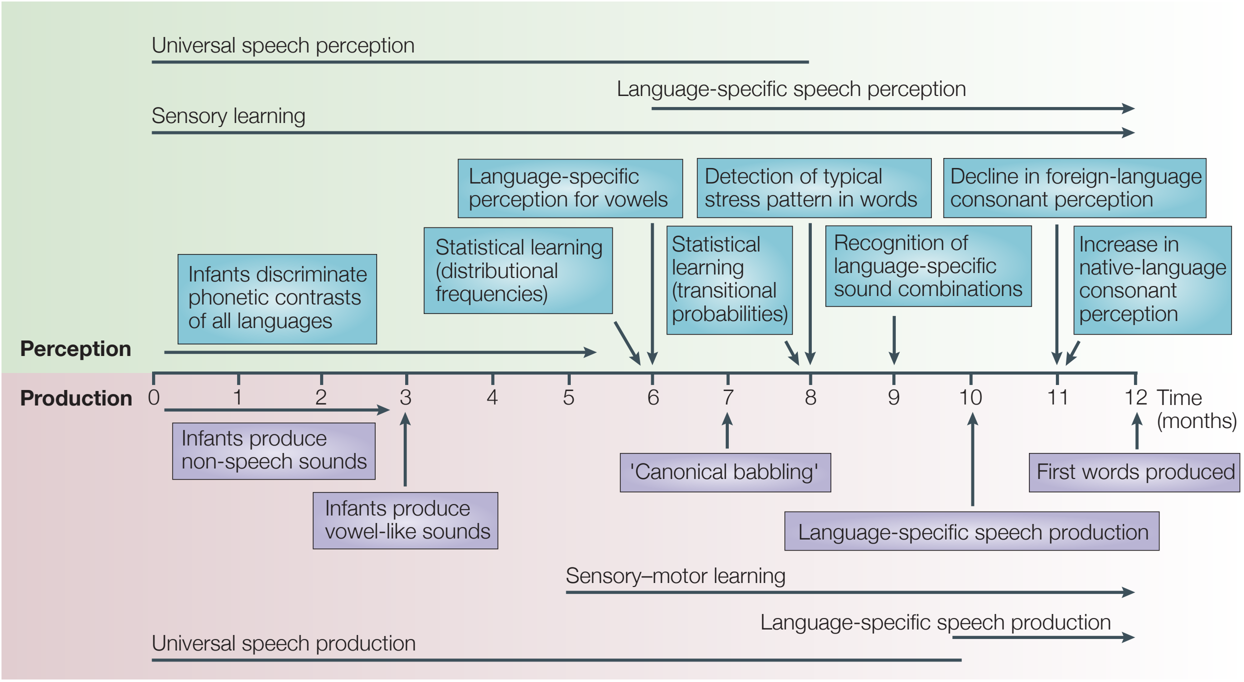
Universal speech perception
Language-specific speech perception
Sensory learning
Language-specific perception for vowels
Detection of typical stress pattern in words
Decline in foreign-language consonant perception
Statistical learning (distributional frequencies)
Statistical learning (transitional probabilities)
Recognition of language-specific sound combinations
Increase in native-language consonant perception
Infants discriminate phonetic contrasts of all languages
Perception
Production
0
1
2
3
4
5
6
7
8
9
10
11
12
Time (months)
Infants produce non-speech sounds
Infants produce vowel-like sounds
'Canonical babbling'
First words produced
Language-specific speech production
Sensory–motor learning
Language-specific speech production
Universal speech production

Figure 1.15: The universal language timeline of speech-perception and speech-production development [Kuhl 2004]. This figure shows the changes that occur in speech perception and production in typically developing human infants during their first year of life.

observed in older children.

In addition to perceptual effects of external environments, such as neonatal imitation, researchers have reported auditory-feedback in toddlers. In adults, a distortion of formant frequencies in auditory-feedback alters online speech production. One study [MacDonald *et al.* 2012], however, reported that real-time formant perturbation does not alter toddlers' speech, suggesting that young children do not have real-time auditory-feedback control.

1.3 QUESTIONS ADDRESSED BY THIS THESIS

As shown in Sec. 1.2, traditional linguistic and acoustic analysis has revealed developmental changes in acoustic phenomena and their consequences. Recent advances in measurement technique have detailed the presence of developmental changes in the anatomy of the speech production system. Nevertheless, the development of motor control in speech production and effects of perceptual cognition on speech production remain open questions. In this thesis, I focused on motor control in speech production and conducted two studies. My aim was to clarify (1) longitudinal changes in vowel articulation as opposed to vowel sounds alone and (2) the divergent process in phonotactics, possibly based on neuromuscular control of articulatory coordination (Fig. 1.16).

In the first study (Chap. 2), I investigate vowel articulation in Japanese children. Because ethical and practical issues prevent articulation investigations in children, I chose a different approach: I estimated articulatory states from acoustical features, in a process known as *acoustic-to-articulatory inversion.* In this study, I showed that vowel articulation went through three stages, and individual analysis potentially supported the result of group

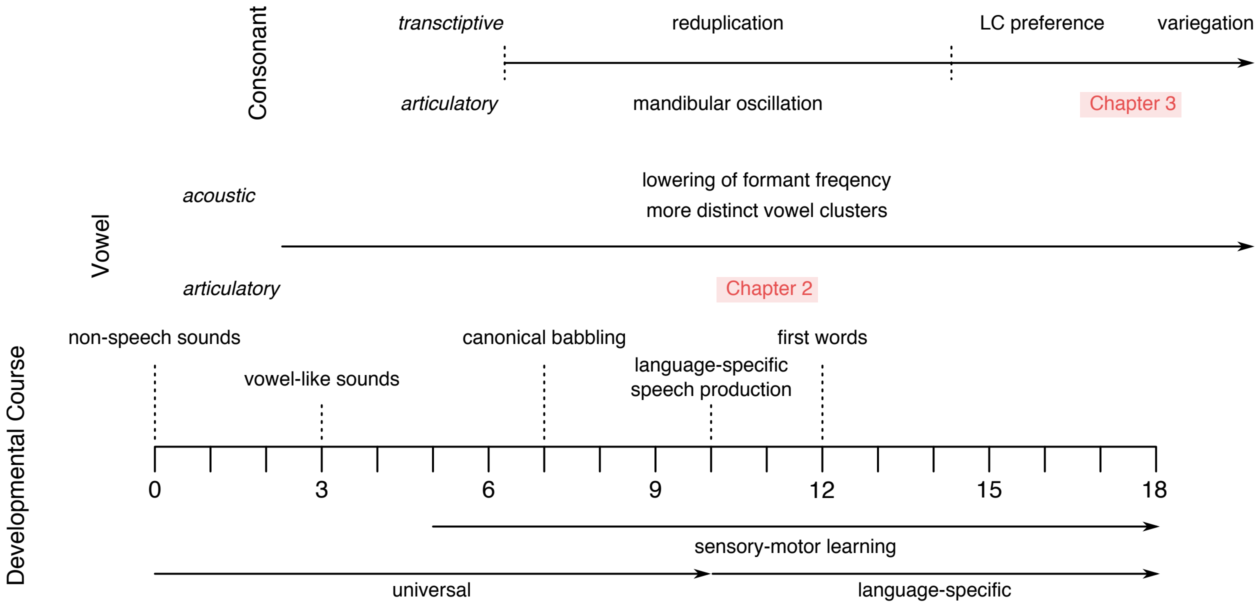

Figure 1.16: Topics dealt with in this thesis in developmental course.

analysis.

In the second study (Chap. 3), I investigated the longitudinal development of consonant-vowel-consonant(-vowel) sequences produced by Japanese and English children. How and when muscle should be activated, thereby moving articulatory organs, are parameters subject to neural works. In this sense, articulatory sequences underlying phonotactics are a neuromuscular problem. From an extreme viewpoint, linguistic structures would be independent of modalities. On the other hand, from the embodiment perspective, linguistic structures are constrained by speech production systems, forcing longitudinal changes in phonological structures to reflect development of neuromuscular coordination of articulatory systems. In this study, I demonstrated that the development of intra- and inter-articulator coordination constrained the acquisition of serial orders in speech with the complexity that characterizes adult language.

Finally, I will summarize the first and second studies, and discuss further issues (Chap. 4).

Chapter 2

Acquisition of Vowel Articulation in Childhood Investigated by Acoustic-to-articulatory Inversion

2.1 INTRODUCTION

The speech sounds are generated by complex motor coordination among the articulatory organs. While the developmental process of speech production has previously been depicted mainly on the basis of evidence derived from acoustical phenomena and their consequences—such as spectral envelope, fundamental frequencies [Amano *et al.* 2006, Ishizuka *et al.* 2007, Kent & Murray 1982, Vorperian & Kent 2007], and phonetic transcriptions [Ingram 1974, MacNeilage & Davis 2000, MacNeialge 2008, Oller 2000, Stoel-Gammon & Cooper 1984] —the development of the articulatory system by which these acoustics are produced still remains an open question because of limitations on the measurement of the articulatory system, especially that of tongue movements. In the present study, I investigated longitudinal changes in children's articulation by estimating the parameters of an articulatory model on the basis of the acoustical features of speech sounds.

The development of speech production during the first year of life has been characterized as following a particular course [Kuhl 2004, Oller 2000, Stoel-Gammon & Cooper 1984]. Infants are born able to produce spontaneous sounds, such as sneezing and crying. Infants then produce cooing, that is, quasivocalic sounds similar to vowels. Subsequently, coos expand into clear vowel sounds characterized by full resonance and wide variety. At an early stage of babbling, infants repeat the same consonant–vowel (CV) units such as "papapa" and "mamama." Finally, beginning around the end of the first year of life, infants produce meaningful speech.

Acoustical studies show that as children grow up, their vowel clusters become more distinct, and the fundamental frequency and spectral peaks (formant frequencies) of their utterances become lower [Amano *et al.* 2006, Ishizuka *et al.* 2007, Kent & Murray 1982, Vorperian & Kent 2007]. Moreover, analyses of phonetic transcriptions show a modification process at work in infants' vocalizations [MacNeilage & Davis 2000, MacNeialge 2008]. At the babbling stage, infants prefer to repeat three predominant CV sequences, that is, labial–central, coronal–front, and dorsal–back CV patterns. With development, children begin to chain variegative CVs, with a fronting tendency in which the first consonant in words has a more anterior place of articulation than the second one [Ingram 1974]. These phenomena are crosslinguistically observed [MacNeialge 2008, Vorperian & Kent 2007].

These changes are likely to be caused mainly by the development of vocal tract anatomy, respiration, and motor controls of articulators. In order to investigate the anatomical structure of the articulatory system and its dynamics during speech production, previous studies have adopted a variety of methods, such as radiographic imaging [Chiba & Kajiyama 1942, Fant 1960, Kiritani 1986], electromagnetic articulography and electropalatography [Byrd & Tan 1996, Hixon 1971], magnetic resonance imaging [Fitch & Giedd 1999, Masaki *et al.* 1999, Vorperian *et al.* 1999, Vorperian *et al.* 2005], ultrasound [Geddes *et al.* 2008, Zharkova *et al.* 2011], and motion-capture systems [Goffman & Smith 1999, Green *et al.* 2000, Green *et al.* 2002, Nip *et al.* 2009]. With regard to anatomy, previous studies reveal that children's vocal tracts, especially during the first year of life, are shaped differently from those of adults [Goldstein 1980, Fitch & Giedd 1999, Sasaki *et al.* 1977, Vorperian *et al.* 1999, Vorperian *et al.* 2005]. Infants' vocal tracts are not only smaller than adults', but they have a broader oral cavity, a tongue mass that is proportionally larger and more anterior, and a more gradually sloping pharyngeal tract. These properties of the infant vocal tract should raise formant frequencies and lead to less clear vowel clusters. In addition, the limited range of tongue movement prevents complex consonantal articulations. While these anatomical changes in vocal tract are certainly responsible for the changes in the filter properties of speech sounds, their phonation is conversely affected mostly by the development of respiration [Boliek *et al.* 1996, Reilly & Moore 2009]. For instance, decrease in the compliance of the chest wall results in more rapid modulation of respiratory muscle movements.

As for the development of motor control of articulators, transcription analysis suggests that infants have relatively independent control over their jaw and that ability to carry out tongue movements depends largely on jaw control [MacNeilage & Davis 2000, MacNeialge

2008]. On the basis of these findings, it has been convincingly argued that mandibular oscillations have a crucial role in the early development of articulation. One study using motion capture partly supports this idea by reporting that jaw movements mature earlier than lip ones [Green *et al.* 2002, Nip *et al.* 2009]. Another study, using electromagnetic articulography and acoustical analysis, reports that fronting tendencies that are predominant in both adults and children are caused by coordination among articulators [Rochet-Capellan *et al.* 2007].

Thus, as described above, the acoustical analysis and empirical measurement of the articulatory system reveals much about the development of speech production. However, many aspects of the development of articulation, including tongue movements crucial to vowel production, especially until the second year of life, still remain an open question. This is because of limitations to the empirical measurement of articulatory movements in young children.

Another approach to investigate articulatory movements is to estimate articulatory states from acoustical features; this is called acoustic-to-articulatory inversion [Atal *et al.* 1978,Hiroya & Honda 2004,Ménard *et al.* 2004,Ouni & Laprie 2005,Shirai 1993,Wakita 1973]. This technique relies on a mapping function from acoustical to articulatory space. Previous studies have proposed several such mapping functions [Atal *et al.* 1978, Hiroya & Honda 2004, Ouni & Laprie 2005, Shirai 1993, Wakita 1973] and, on their basis, articulatory models [Maeda 1990, Mermelstein 1973, Story 2009]. When it comes to applying this technique to sounds produced by infants, however, some problems arise. First, because of anatomical differences between infants' vocal tracts and those of adults, the articulatory model used must be scalable to the child's vocal tract size. Second, I cannot calculate a mapping function from acoustical to articulatory features, since it is impossible to pair acoustical features with empirically obtained articulatory features in this case. Third, the model, which has enough, but not too many parameters to approximate the vocal tract shape, is appropriate.

Taking into consideration the need for scalability of the vocal tract and parameters to specify articulatory states, I adopted Maeda's model [Maeda 1990, Ménard *et al.* 2004, Serkhane *et al.* 2007]. Originally, this model was proposed to approximate midsagittal slices of the vocal tract during adult' vowel productions [Maeda 1990]. In the model, seven parameters are retrieved by a factor analysis of vocal tract contours and associated with the articulatory organs and their parameters: the jaw, tongue dorsum position, tongue dorsum shape, tongue apex position, lip aperture, lip protrusion and larynx height (Fig. 2.1A). Each parameter is adjustable to values in the range of ±3.5 standard deviations around the mean, and a linear sum of parameters generates the midsagittal contour of the vocal tract.

Subsequent studies [Ménard *et al.* 2004, Serkhane *et al.* 2007] propose two scaling factors to incorporate growth data [Goldstein 1980] into the model and apply it to non-adult-sized vocal tracts. The two scaling factors, which respectively control the length of the oral and pharyngeal cavities, enable us to apply the model to different sizes of vocal tract (Fig. 2.1B). On the semipolar grids (in the right panel of Fig. 2.1B), the 5mm intergrid spaces, as well as the midsagittal distance between the ventral and dorsal contours of the vocal tract, were scaled by one of the two factors or by an interpolation of them. The values of the factors were calibrated month by month, on the basis of the previous studies [Goldstein 1980, Ménard *et al.* 2004, Serkhane *et al.* 2007] (the right panel of Fig. 2.1B). By

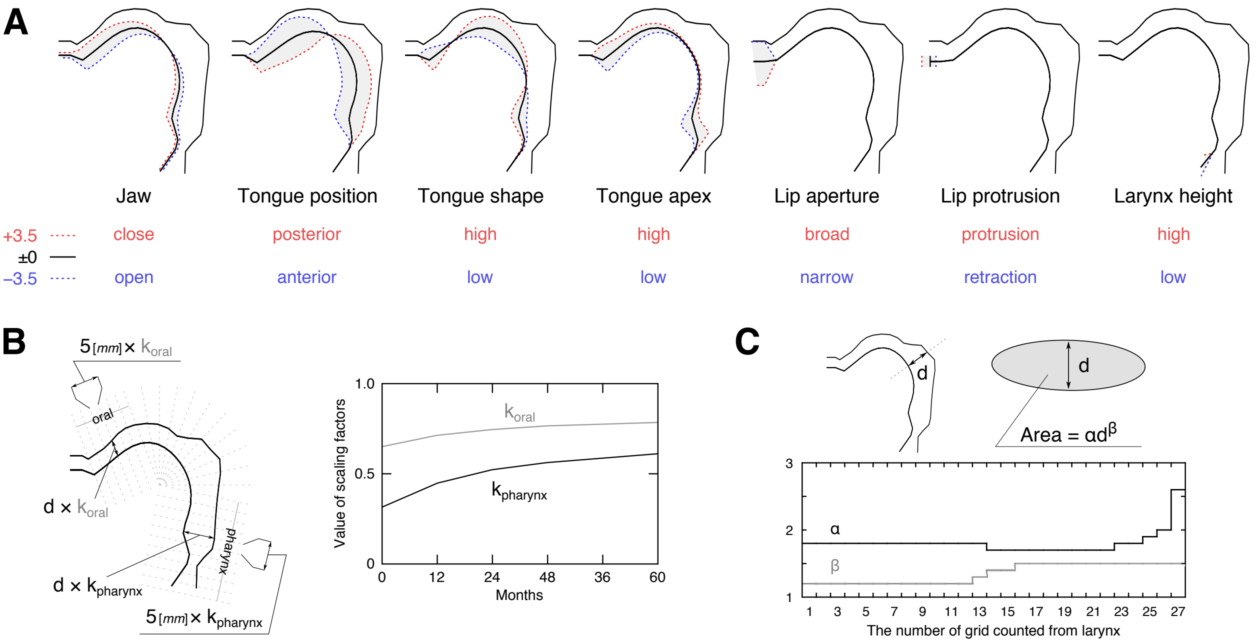

Figure 2.1: The parameters and scaling factors of the articulatory model. (A) The seven articulatory parameters and their movements. Solid and dashed lines indicate the shape of the vocal tract when the parameter values are zero, ±3.5, respectively. (B) The schematic representation of two scaling factors for vocal tract length and their values as a function of age (in months). (C) The calculation of the vocal tract area function by α-β transformation.

comparing simulated formant frequencies obtained using the scalable Maeda's model with actual formant frequencies produced by infants at 4 and 7 months of age, a previous study [Serkhane *et al.* 2007] argues that the jaw plays only a minor role before the babbling stage but a major role at the onset of rhythmic syllable-like output in canonical babbling. This argument is based on the assumption that infants can adopt a full range of parameter values only for tongue position, tongue shape, jaw and lip height. There are, however, two problems with this assumption: First, if infants can take only a limited rage of parameter values, the analysis will overestimate possible range of the space of speaker's formant frequencies. Second, although the model focuses on the four articulators, the involvement of the other articulators should not be a priori excluded. Thus, in this study, in order to analyze the articulatory development of vowel production in Japanese children, I performed an acoustic-to-articulatory inversion technique using the scalable Maeda's model, with seven articulatory parameters. For materials, I used the vowel sounds of Japanese, which consist of high-front /i/, mid-front /e/, low-center /a/, high-back /u/ and mid-back /o/, as produced by three children over time from ages 6 to 60 months. These data were taken from the NTT infant database [Amano *et al.* 2006, Amano *et al.* 2009, Ishizuka *et al.* 2007].

2.2 METHODS

2.2.1 Materials

As mentioned above, I used the NTT Japanese infant speech database [Amano *et al.* 2006, Amano *et al.* 2009, Ishizuka *et al.* 2007] for this study. This database contains the utterances of five normally developing children and their parents, recorded with 8-bit quantization at a sampling rate of 16 kHz. Two well-trained transcribers segmented and labeled the speech data in terms of the Japanese phoneme inventory. The phonetic transcriptions were double-checked. In this study, I analyzed the utterances of three children (one boy and two girls: hereafter, I refer them as child B, child C and child D, respectively). Because of lacks of transcriptions, I did not use data of the other two children. To exclude ambiguous utterances, I used only phoneme labels that were agreed on by the two transcribers. Furthermore, I excluded utterances that included crying, laughing, or any other sound than speech. I randomly selected 60 samples of each child per each Japanese vowel (/i/, /e/, /a/, /u/ and /o/) for each month of age (3 children × 60 samples × 5 vowels × 6–60 months). Note that there are many instances in which 60 samples are not available at younger ages. Thus, the number of data used for the analysis differed among months of age.

2.2.2 Acoustic parameter estimation

I obtained formant frequencies as filter characteristics, using the linear predictive coding (LPC) analysis method [Atal & Hanauer 1971, Itakura 1975]. I present a peseudo-code of LPC analysis in Alg. 1. To as far as possible avoid coarticulation effects, I analyzed 25ms that began at the first quarter position of the total duration of each labeled vowel. The samples were pre-emphasized to equalize energy across the spectrum and improve spectral model fitting; (the pre-emphasis coefficient was 0.97); then, I applied LPC analysis. I set 12 as the order of LPC analysis, since I confirmed that the extraction of the three lowest formant frequencies (F_1, F_2, and F_3) by 12$^{\text{th}}$-order LPC analysis maximized the correct classification rate, obtained by linear discriminant analysis with the vowel categories as the predictor and the three formants mentioned as independent variables. F_1, F_2, and F_3 were automatically extracted by peak-picking spectrum envelopes. The property of children's speech in which fundamental frequency (F_0) is higher than that of adults prevented the extraction of precise values for the formant frequencies [Vallabha & Tuller 2002]; thus, I eliminated data whose estimated F_1 fell into the range of ±10% of the F_0. I used F_0 values already provided in the database. These data were estimated using a robust F_0 estimation method [Nakatani & Irino 2004] and subsequently manually corrected. In order to evaluate the accuracy of the formant estimation method used in this study, I calculated estimation errors, in the following manner. First, I calculated theoretical F_1, F_2 and F_3, and their bandwidths using Maeda's model with neutral parameters (that is, all seven parameters were zero) and scaling factors adjusted from 6 to 18 months; this will be further explained in a later subsection. Second, based on these formant frequencies and bandwidths, I obtained a time domain signal using the Klatt cascade synthesizer [Klatt 1980]. At this time, we set F_0 to a value ranged from 200 to 500 Hz by 50 Hz step. Finally, I estimated formant frequencies from the signal, whose F_0 were ranged from 200 to 500 Hz. In this manner, I calculated errors in formant estimation over different fundamental frequencies and different months of age. The estimation error was within around 20% of the fundamental frequencies at all months we conducted.

Note that previous studies using the NTT infant database have already shown developmental changes in fundamental and formant frequencies [Amano *et al.* 2006, Ishizuka *et al.* 2007]. However, the process of development of articulatory control in children is not shown. Therefore, I attempted to estimate articulatory control on the basis of the acoustical features described in the following subsection.

2.2.3 Articulatory synthesizer

As I describe in Sec. 2.2.4, in order to estimate articulatory feature, it is indispensable to model the vocal tract and to calculate a transfer function corresponding to the vocal tract. Here, I introduce the way to model the vocal tract and to calculate a transfer function. To do so, I went through three steps; (i) transformation from articulatory parameters to a midsagittal shape of the vocal tract, (ii) transformation from a midsagittal shape of the vocal tract to an area function, and (iii) calculation a transfer function based on an area function.

Algorithm 1 The pseudo-code of the linear predictive coding is as follows. In the pseudo-code, a normal (*e.g.*, x) and bold symbol (*e.g.*, $\mathbf{y}$) represent scalar and vector, respectively.

```
INPUT: Input signal $\mathbf{s} = \{s_1, \cdots, s_n\}$, prediction order $m$, a coefficient for pre-emphasis $\alpha$
OUTPUT: LPC coefficient $\mathbf{a} = \{a_i \cdots a_m\}$
Pre-emphasis input signal to exclude a property of radiation: $s_t \leftarrow s_t - \alpha s_{t+1}$
Multipling pre-emphasised signal by a window function $\mathbf{w}$ such as hanning window: $s_t \leftarrow s_t - w_t s_t$
Calculation for autocorrelation $\mathbf{v} = \{v_i\}$: $v_i = \frac{1}{N}\sum_{t=i}^{N-i} s_t s_{t-i}$ under the condition $v_{-i} = v_i \ (i = 0 \cdots m)$
Multipling $\mathbf{v}$ by lag window $\mathbf{h}$: $v_i \leftarrow h_i v_i$
Solve Yule-Walker equation to obtain $\mathbf{a}$ by Levinson-Durbin-Itakura algorithm by follow loop. Initially, set residual power $u_0 = v_0$.
for $i = 1$ to $m$ do
    Calculation for PARCOR coefficient $k_i$: $k_i = \sum_{j=1}^{i-1}\{a_j^{(i-1)} v_{i-j} + v_i\}/u_{i-1}$
    $a_i^{(i)} = -k_i$
    for $j = 1$ to $i - 1$ do
        $a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}$
    end for
    Calculation for residual power $u_i$: $u_i = (1 - k^2)u_{i-1}$
end for
```

I explain transformation from articulatory parameters to a midsagittal shape of the vocal tract. As I mentioned in Sec. 2.1, previous studies propose several articulatory models and, in the present study, I adopted Maeda's model, whose seven parameters adjustable to values in the range of ±3.5 standard deviations around the mean. The seven articulatory parameters associated with the articulatory organs are shown in Fig. 2.1A. Correspondences between articulatory-parameter value and states of the articulatory system are shown in

Fig. 2.1A: positive values of the parameters for the jaw, tongue position, tongue shape, tongue apex, lip aperture, lip protrusion and larynx height respectively represent narrow vocal tract, posterior tongue position, high tongue apex, high tongue dorsum, broad lip opening, protruded lips and high larynx position. Note that the parameter for tongue apex position seems to affect the tongue contour in the laryngeal region as well as in the tongue apex region. A previous study [Maeda 1990] speculates that this apparent effect of the tongue apex component in the laryngeal region is due to orthogonality among the components, imposed by the statistical method. Putting each contour of the i$^{\mathrm{th}}$ articulatory parameters as $\boldsymbol{x}^{\mathbf{(i)}} = \{x_1^{(\mathrm{i})}, \cdots, x_n^{(\mathrm{i})}\}$ where n is the number of section of the vocal tract, and weight matrix of the articulatory parameters as $\boldsymbol{w} = \{w_1, \cdots, w_7\}^{\mathrm{T}}$, then the midsagittal contour of the vocal tract is obtained by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$ where $\boldsymbol{X} = \{{\boldsymbol{x}^{\mathbf{(1)}}}^{\mathrm{T}}, \cdots, {\boldsymbol{x}^{\mathbf{(7)}}}^{\mathrm{T}}\}$.

Next, I transformed from a midsagittal shape of the vocal tract to an area function by α-β transformation. Giving a midsagittal contour of the vocal tract, I obtained a distance from an ventral contour to an dorsal contour at each grid point (Fig. 2.1B). Subsequently, I multiplied the distances by scaling factors, which are shown in Fig. 2.1B. Then, I calculated areas at each sections A_i by an equation $A_i = \alpha_i d_i^{\beta_i}$ where d_i, α_i and β_i are the distance at i$^{\mathrm{th}}$ section and coefficients shown in Fig. 2.1C, respectively.

Finally, I calculated a transfer function based on the area function by the acoustic tube model. Roughly speaking, there are two way to calculate the frequency: calculations in the frequency [Fant 1960, Flanagan 1972, Sondhi & Schroeter 1987] and time domain [Ho *et al.* 2011, Maeda 1982, Mokhtari *et al.* 2008]. In the present study, I calculated a transfer function in time domain proposed in Flanagan (1972) [Flanagan 1972] as follows. First, I introduce the equation of continuity. As I described in the previous section, the vocal tract shape can be modeled by a sequence of cylinders. Putting a length and area of cylinder as δx and A, respectively, volume of cylinder can be written by $W_0 = A\delta x$. Moreover, their change can be written as follows:

$$\begin{aligned} w(x,t) &= A(\xi(x+\delta x,t) - \xi(x,t)) \\ &= W_0 \frac{\partial \xi(x,t)}{\partial x}, \end{aligned} \tag{2.1}$$

where $\xi(x,t)$ denotes that particle at position x moves $\xi(x,t)$ at time t. On the other hand, pressure differences p can be written as follows:

$$p = -\eta P_0 \frac{w}{W_0} = -K \frac{w}{W_0}, \tag{2.2}$$

where w denotes volume differences, $K = \eta P_0$ is bulk modulus, and P_0 is a pressure. Combining Eqs. 2.1 and 2.2, and then differentiating it in terms of time, we can obtain the following equations called as the equation of continuity:

$$p(x,t) = -K \frac{\partial v(x,t)}{\partial x}, \tag{2.3}$$

where $v(x,t) = \partial \xi(x,t)/\partial t$ is particle velocity. Second, I introduce the equation of motion. Putting ρ as the density of air, the mass of air within the volume W_0 is $W_0\rho = A\delta x\rho$ and we obtain the equation of motion by differentiating the particle velocity $v(x,t)$ in terms of time as follows:

$$Ap(x,t) - Ap(x+\delta x,t) = A\delta x\rho \frac{\partial v(x,t)}{\partial t}. \tag{2.4}$$

Setting $\delta x = 0$, we obtain

$$\frac{\partial p(x,t)}{\partial x} = -\rho \frac{\partial v(x,t)}{\partial t}. \tag{2.5}$$

Next, I introduce the wave equation. Putting x, $A(x)$, $p(x,t)$, $v(x,t)$ and $u(x,t)$ as distance from glottis along with a midline of the vocal tract, area function, air pressure, particle velocity and volume velocity, respectively, and considering $u(x,t) = A(x)v(x,t)$, we obtain the following equation to combine the equations of continuity and motion:

$$\left\{ \begin{array}{lcl} \frac{\partial u(x,t)}{\partial x} & = & -\frac{A(x)}{K}\frac{\partial p(x,t)}{\partial t} \\ \frac{\partial p(x,t)}{\partial x} & = & -\frac{\rho}{A(x)}\frac{\partial u(x,t)}{\partial t}. \end{array} \right. \tag{2.6}$$

By eliminating $p(x,t)$ or $u(x,t)$, we obtain the wave equations:

$$\left\{ \begin{array}{lcl} \frac{A(x)}{c^2}\frac{\partial p(x,t)}{\partial t^2} & = & \frac{\partial}{\partial x}\left(A(x)\frac{\partial p(x,t)}{\partial x}\right) \\ \frac{1}{c^2 A(x)}\frac{\partial u(x,t)}{\partial t^2} & = & \frac{\partial}{\partial x}\left(\frac{1}{A(x)}\frac{\partial u(x,t)}{\partial x}\right). \end{array} \right. \tag{2.7}$$

When e^{jwt} represents the time factor, discretized forms of the wave equation in a cylinder of which length is l are as follows:

$$\left\{ \begin{array}{lcl} \frac{\mathrm{d}U(x)}{\mathrm{d}x} & = & -j\omega\frac{A}{K}P(x) \\ \frac{\mathrm{d}P(x)}{\mathrm{d}x} & = & -j\omega\frac{\rho}{A}U(x), \end{array} \right. \tag{2.8}$$

where $U(x)$ and $P(x)$ are distributions of volume velocity and air pressure. On the other hand, the telegraphy equation is follows:

$$\left\{ \begin{array}{lcl} \frac{\mathrm{d}I(x)}{\mathrm{d}x} & = & -(G + j\omega C)E(x) \\ \frac{\mathrm{d}E(x)}{\mathrm{d}x} & = & -(R + j\omega L)I(x), \end{array} \right. \tag{2.9}$$

where R, L, G, C, $E(x)$ and $I(x)$ denote a serial resistance, serial inductance, parallel conductance, parallel capacitance, current and electric pressure, respectively. Assuming that

$$R = G = 0, \quad L = \frac{\rho}{A} \text{ and } C = \frac{A}{K} = \frac{A}{\rho c^2}, \tag{2.10}$$

the transmission line is equivalent to the wave equation. But, in fact wall of the vocal tract is yielding. So,

$$\begin{aligned} R &= \frac{S}{A^2}\sqrt{\frac{\omega\rho\mu}{2}} \quad \text{and} && (2.11) \\ G &= S\frac{\eta-1}{\rho c^2}\sqrt{\frac{\lambda\omega}{2C_p\rho}}, && (2.12) \end{aligned}$$

where S, μ, λ, C_p and η denote a perimeter of a cylinder, viscous modulus, heat conductivity, constant pressure specific heat and ratio of constant pressure specific heat to specific heat at constant volume, respectively. The solutions of the transmission line equivalent to wave equation can be written as follows:

$$\begin{cases} P(x) &= Z_C(U^+e^{-\gamma x} + U^-e^{-\gamma x}) \\ U(x) &= U^+e^{-\gamma x} + U^-e^{-\gamma x} \end{cases} \tag{2.13}$$

where

$$\begin{aligned} Z_C &= \sqrt{\frac{R+j\omega L}{G+j\omega C+Y_w}} \quad \text{and} && (2.14) \\ \gamma &= \sqrt{(R+j\omega L)(G+j\omega C+Y_w)}. && (2.15) \end{aligned}$$

In the above equation, P^+ and U^+ denote each component in forward-traveling wave, and P^- and U^- denote each component in backward-traveling wave. Eliminating U^+ and U^- results in

$$\begin{pmatrix} P(0) \\ U(0) \end{pmatrix} = \begin{pmatrix} \cosh(\gamma l) & Z_C\sinh(\gamma l) \\ \frac{1}{Z_C}\sinh(\gamma l) & \cosh(\gamma l) \end{pmatrix} \begin{pmatrix} P(l) \\ U(l) \end{pmatrix}. \tag{2.16}$$

Since

$$\left.\begin{aligned} P_i(l) &= P_{i+1}(0) \\ U_i(l) &= U_{i+1}(0) \end{aligned}\right\}, \tag{2.17}$$

$$\begin{pmatrix} P_g \\ U_g \end{pmatrix} = \boldsymbol{F_1 F_2 \cdots F_n} \begin{pmatrix} P_l \\ U_l \end{pmatrix}. \tag{2.18}$$

Putting

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \boldsymbol{F_1 F_2 \cdots F_n}, \tag{2.19}$$

frequency responses are given as follows:

$$Z_{\text{in}} = \frac{AZ_r + B}{CZ_r + D}, \tag{2.20}$$

where Z_r is radiation impedance.

2.2.4 Articulatory state estimation

In order to estimate a set of the articulatory parameters of Maeda's model from formant frequencies, I conducted iterative optimization with an inversion procedure exploiting the pseudo-inverse Jacobian matrix [Atal *et al.* 1978, Ménard *et al.* 2004]. Note that this is an attempt to solve an ill-posed problem. To address this, I repeatedly performed iterative optimization with random initializations. Concretely, I performed the following process thirty times per sample to estimate articulatory parameters (Fig. 2.2): (i) Following the previous studies [Goldstein 1980, Ménard *et al.* 2004], I set values for scaling factors of specific months. (ii) I set random values into $\boldsymbol{p} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$, where $p_1, p_2, p_3, p_4, p_5, p_6$ and p_7 denote the jaw, tongue position, tongue shape, tongue apex, lip aperture, lip protrusion and larynx height parameters, respectively, as initial parameters. (iii) In order to obtain pseudo-inverse Jacobian matrix $\boldsymbol{J}^+$, I calculated the Jacobian matrix by $\boldsymbol{J} = [(\boldsymbol{\delta F_i/\delta p_1}) \cdots (\boldsymbol{\delta F_i/\delta p_7})]$, where $\boldsymbol{\delta F_i/\delta p_j} = (\delta F_1/\delta p_j, \cdots, \delta F_3/\delta p_j)^{\mathrm{T}}$ represents

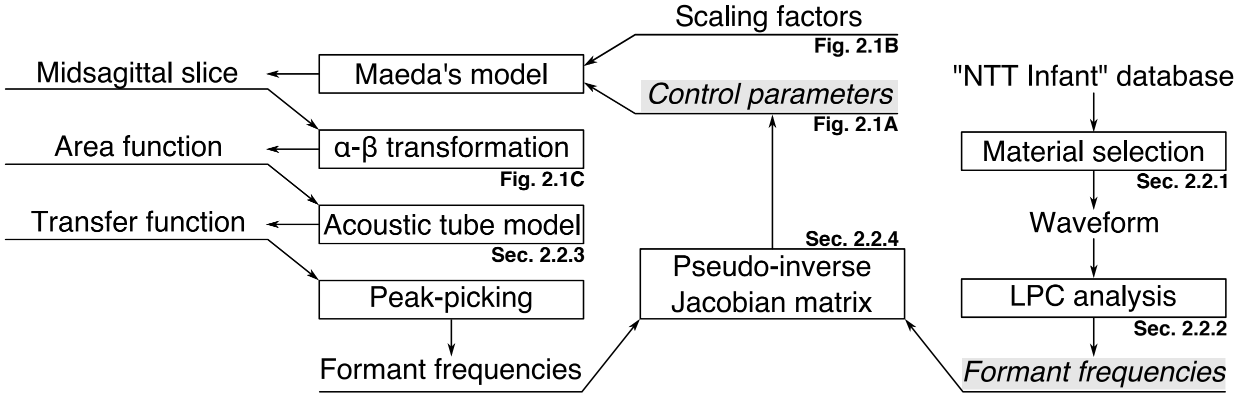

Figure 2.2: The process of estimation of articulatory parameters from acoustic parameters. Iterative optimization was conducted through the forward transformation of control parameters to formant frequencies based on the Maeda's model, and inverse transformation of formant frequencies to control parameters was conducted using a pseudo-inverse Jacobian matrix. For details, see Sec. 2.2.4.

the vector of the partial derivation of F_1, F_2 and F_3 formant frequencie with respect to the j^{th} articulatory parameter. Then I applied singular value decomposition to the Jacobian matrix. (iv) I calculated formant frequencies $\boldsymbol{a} = \{F_1, F_2, F_3\}$ based on the set of articulatory parameters $\boldsymbol{p}$ in the manner described in the above paragraph. (v) I updated the set of articulatory parameters $\boldsymbol{p}$ by calculating $\boldsymbol{p} \leftarrow \boldsymbol{p} + \eta \times \boldsymbol{J}^{+} \times (\boldsymbol{a_0} - \boldsymbol{a})$, where $\boldsymbol{a_0}$ is the formant frequencies of the vowel measured by LPC analysis and η is the learning rate. (vi) I calculated an optimized set of articulatory parameters by repeating (iii)–(v) thirty times (Thus, I conducted the optimization process 900 times = 30 initial values × 30 iterations). If any estimated parameters were larger or smaller than ±3.0, I discarded the set.

2.2.5 Validation of the inversion

In nature, I would ideally compare vocal tract shapes estimated by the inversion technique and those measured by empirical techniques such as magnetic resonance imaging. However, since measurement of the vocal tract and articulatory parameters during speech articulation is difficult in children, in this study I instead used three other methods to evaluate the validation of the acoustic-to-articulatory inversion technique.

First, to evaluate convergence of inversion calculation through iterative optimizations, I calculated averages of absolute differences between formant frequencies extracted from vowel sounds produced by children and the inversely estimated formant frequencies.

Second, I attempted to evaluate the validation of the inversion technique by comparing measured area functions with inversely estimated ones. However, since the actual area function of the infants' vocal tract was not available, I conducted the comparison by using area functions for adults. The area functions corresponding to the five vowels were derived from adult functions using VTCalcs [http://www.cns.bu.edu/ speech/VTCalcs.php]. Based on these area functions, I conducted forward-transformation to obtain formant frequencies, as follows. /i/: (F_1, F_2, F_3) = (215, 2012, 3324), /e/: (349, 1895, 2723), /a/: (605, 1267, 2066), /u/: (244, 801, 2005), and /o/: (435, 806, 1817) [all formant figures are in Hz]. Subsequently, I inversely estimated area functions from the forward-transformed formant frequencies, and finally, I compared the original area functions with the inversely estimated ones.

Third, I evaluated the stability of the inversion technique for the estimation of articulatory parameters, as follows. (i) Since actual values are not available for the articulatory parameters in infancy, I calculated a set of mean values for the inversely estimated articulatory parameters for each vowel at 12 months. (ii) I calculated formant frequencies using these values. (iii) I applied the inversion technique to the formant frequencies and obtained value estimates for the articulatory parameters. (iv) I repeated (ii) and (iii) 60 times each for vowel and obtained mean values for the inversely estimated articulatory parameters. (v) Eventually, I compared the original mean value for each articulatory parameter with the inversely estimated mean value for the same parameter. If forward and inverse transformation provides an exact solution, the difference between the original and transformed values will be close to zero.

2.2.6 Analysis of developmental changes in articulatory parameters

In order to investigate longitudinal changes in the articulatory parameters over the children's development, I calculated the mean values of the articulatory parameters. Because of the small number of samples per month, data for two adjacent months were merged for data from when the children were less than 30 months of age. Then, to reveal the trend of the mean values over development, I applied the third-order polynomial function to the values with months of age as the independent variable. The number of inversely estimated samples of each month is shown in Table 2.1 (the *total* columns). For instance, I pooled data at 6 and 7 months in Table 2.1, and regarded a set of data as data at 6–7 months. Because of the small number of samples in early development, the number of successfully inverse-estimated parameters tends to be smaller for earlier months of age.

2.2.7 Analysis of distinctness

Many previous acoustical studies have focused on distinctness among the formant frequencies for each vowel in children's speech [Ishizuka *et al.* 2007, Vorperian & Kent 2007]. In the present study, in contrast, I evaluated distinctness among the articulatory parameters of each vowel. The aim of this analysis was to quantify the degree of differentiation of articulatory states during vowel production.

I conducted linear discriminant analysis (LDA) with the vowel categories attached by the transcribers as predictors and the inversely estimated articulatory parameters as independent variables. I selected significant independent variables in a stepwise manner (p <0.05) and recruited them into the classification function. I obtained a correct classification rate calculated based on a confusion matrix for LDA. As the analysis of developmental changes in articulatory parameters, data for two adjacent months were merged for data from when the children were less than 30 months of age. Therefore, the number of samples used for the LDA is equivalent to those for the analysis of developmental changes in articulatory parameters (the *total* columns in Table 2.1). A larger value of the rate indicates that children produce each vowel using more distinct articulatory states. Note that significance only indicates that the variable contributes to differential production of the vowel: if a variable contributes equally to the articulation of all the vowels, the variable will not show significance. I also obtained confusion matrices for the acoustic space; again, I calculated F_1, F_2, and F_3 by 12th-order LPC analysis, and then conducted LDA in a stepwise manner (p <0.05). Regarding distinctness in the articulatory space, I further performed individual analysis to examine inter-child variability. In this analysis, because of the small number of samples per month

Table 2.1: The number used for linear discriminant analysis (LDA).

Month	Child B					Child C					Child D					Total				
	/i/	/e/	/a/	/u/	/o/	/i/	/e/	/a/	/u/	/o/	/i/	/e/	/a/	/u/	/o/	/i/	/e/	/a/	/u/	/o/
6	-	-	-	-	-	-	2	3	2	1	20	35	39	39	1	20	37	42	41	2
7	-	-	-	-	-	4	11	34	15	1	-	-	-	-	-	4	11	34	15	1
8	2	26	9	25	1	4	6	6	6	-	2	6	8	5	-	8	38	23	36	1
9	9	36	16	16	-	23	43	42	48	2	6	15	29	20	1	38	94	87	84	3
10	8	18	16	4	-	-	-	2	4	-	49	55	50	53	10	57	73	68	61	10
11	5	16	23	18	-	1	2	5	7	1	13	11	30	17	4	19	29	58	42	5
12	23	46	48	44	1	8	10	36	13	2	43	53	53	56	17	74	109	137	113	20
13	3	24	25	35	3	1	3	4	6	-	3	1	19	5	1	7	28	48	46	4
14	12	44	47	7	7	51	42	51	59	49	2	-	21	11	1	65	86	119	77	57
15	16	34	39	34	5	36	59	51	42	16	1	7	18	18	2	53	100	108	94	23
16	4	20	26	7	14	58	38	51	14	35	33	50	52	50	28	95	108	129	71	77
17	18	39	46	7	27	58	53	53	53	50	5	-	10	2	5	81	92	109	62	82
18	-	-	-	-	-	58	47	56	57	57	36	24	57	37	55	94	71	113	94	112
19	60	28	50	31	33	58	57	55	56	53	31	8	50	16	19	149	93	155	103	105
20	-	-	-	-	-	60	58	54	56	54	60	60	55	60	60	120	118	109	116	114
21	53	56	49	51	46	55	56	43	53	58	-	-	-	-	-	108	112	92	104	104
22	-	-	-	-	-	58	55	57	58	59	37	33	57	32	60	95	88	114	90	119
24	60	60	42	52	53	60	59	59	56	60	60	60	56	58	54	180	179	157	166	167
25	-	-	-	-	-	56	60	54	57	58	57	60	59	55	60	113	120	113	112	118
30	13	3	19	18	24	58	51	57	45	58	56	60	59	58	48	127	114	135	121	130
34	42	48	47	31	49	-	-	-	-	-	51	34	39	39	35	93	82	86	70	84
35	-	-	-	-	-	58	58	57	58	58	18	3	10	2	7	76	61	67	60	65
40	58	60	53	52	54	58	60	53	47	48	49	50	57	39	51	165	170	163	138	153
44	-	-	-	-	-	-	-	-	-	-	37	55	60	41	43	37	55	60	41	43
45	59	56	47	57	42	58	60	53	50	50	18	21	51	5	10	135	137	151	112	102
50	-	-	-	-	-	56	57	54	49	53	-	-	-	-	-	56	57	54	49	53
52	57	60	36	51	43	-	-	-	-	-	60	60	58	60	57	117	120	94	111	100
55	-	-	-	-	-	59	60	55	32	55	-	-	-	-	-	59	60	55	32	55
60	-	-	-	-	-	53	59	49	40	53	-	-	-	-	-	53	59	49	40	53

per child, data for four adjacent months were merged. The numbers of samples used in the individual analysis are shown in Table 2.1 (the *child B*, *child C* and *child D* columns).

Moreover, in order to evaluate the sharpness of articulatory clusters, I obtained logarithmic values of likelihood ratios generated by the classification function. LDA assigns each sample a probability of being classified into each vowel. Thus, the following value L of samples correctly classified indicates the sharpness of articulatory clusters: $L = \log(P_{\text{correct cluster}} / \langle P_{i \in \{\text{incorrect clusters}\}} \rangle)$, where P_x is the probability of being classified into a cluster x and $\langle \boldsymbol{x} \rangle$ denotes the mean of numbers in $\boldsymbol{x}$. A larger L means that the probability of being classified into a correct vowel cluster is higher and the probability of being classified into an incorrect vowel cluster is lower. For example, if $L = \log(3)$, samples will be three times more likely to be classified into a correct than an incorrect cluster.

Using the above method, I conducted classification based on the supervised information on the vowel labels. In other words, I attempted to identify a set of significant parameters contributing to the classification of vowel labels produced by well-trained transcribers. Aside from these attempts, I also tried to estimate articulatory clusters in an unsupervised manner. Concretely, I first, for a dimensional compression, obtained three independent variables by applying independent component analysis, using a fixed-point algorithm, to inversely estimated articulatory parameters, which pooled in each of the three stages (6–9, 10–17 and 18–months in the total columns in Table 2.1). Subsequently, I conducted variational Bayesian inference for Gaussian model mixture (VB) [Bishop 2007] to classify the independent variables.

To quantify the properties of articulatory clusters, I calculated three indices to assess how much all clusters occupied the overall articulatory states (share of clusters), whether each cluster is consistent with a single vowel category (purity within a cluster), and whether a set of clusters explains all of a vowel categories (completeness of vowels). To determine share of clusters, I calculated the Boltzman–Shannon entropy of cluster frequency. For example, if 10 samples were classified into three clusters and the number of samples in each cluster was (c_1, c_2, c_3) = (6, 3, 1), where c_i denotes the number of a cluster, then the index could be calculated by $-\sum_{i=1}^{3}[c_i/10 \times \log_3(c_i/10)]$. For purity within a cluster, first, I calculated the share of vowels within each cluster. Then, I obtained the index by dividing maximum share by mean share of the rest. For example, if ten samples joined a cluster and the number of labels of the samples were as follows: (/i/, /e/, /a/, /u/, /o/) = (5, 2, 1, 1, 1), then the index could be calculated by $5/\langle 2, 1, 1, 1\rangle$. For completeness of vowels, I first obtained vowel categories corresponding to a maximum share within each cluster. Subsequently, I calculated the Boltzman–Shannon entropy of their frequencies to base five (the number of Japanese vowels). For example, if VB suggested ten clusters and the frequency of arguments of the maximum share of each vowel within each cluster were as follows: (/i/, /e/, /a/, /u/, /o/) = (2, 2, 4, 1, 1), then the index could be calculated by $-\sum_{x\in\boldsymbol{v}}[x/10 \times \log_5(x/10)]$, where $\boldsymbol{v}$ = {/i/, /e/, /a/, /u/, /o/}.

2.3 RESULTS

2.3.1 Validation of inversely estimated articulatory parameters

The averages and standard deviations of differences between formant frequencies extracted from vowel sounds and those forward-transformed by the inversely estimated area function from the formant frequencies were as follows: (mean ± 1S.D.); F_1: 8.4 ± 20.2 Hz, F_2: 9.4 ± 15.1 Hz, and F_3: 15.0 ± 30.6 Hz.

I also evaluated the inversion technique based on the area functions. Figure 2.3 suggests that the area functions generating formant frequencies were similar to the inversely estimated ones from the formant frequencies (means ± 1 S.D.s of error were as follows: /i/: 0.58 ± 0.51 cm^2, /e/: 0.72 ± 0.97 cm^2, /a/: 1.53 ± 2.13 cm^2, /u/: 1.11 ± 0.59 cm^2, and /o/: 0.80 ± 0.94 cm^2). Since the lip protrusion and larynx height parameters contribute to the length of the vocal tract along a centerline of the vocal tract (Fig. 2.1A), there were mismatches between original vocal tract lengths and those of the inversely estimated vocal tract. Thus, I truncated unoverlapped area in the error calculations. Although some estimation errors existed, the places of constriction and release in the vocal tract were consistent with our standard, that is, with the assumption that I make anterior constrictions of the vocal tract to articulate /i/ or /e/, whereas I open the anterior of the vocal tract to produce /a/ or /o/. In addition, I open the lips to produce /a/, yet narrow them to produce /u/ and /o/.

Furthermore, to evaluate the stability of the inversion technique for the estimatation of articulatory parameters, I examined whether forward and inverse transformations of the set of articulatory parameters for each vowel, which were actually chosen from values estimated based on sound data for 12-month-olds, produce differences between transformed and original values. As shown in Table 2.2, differences in the parameters, with the exception of lip aperture, ranged from 0.001 to 0.657 (0.015 to 11% of parameter codomain). Differences in the lip aperture parameter were around 1.0 (17% of parameter codomain). These findings support the validity of the inversion technique for the estimation of articulatory parameters.

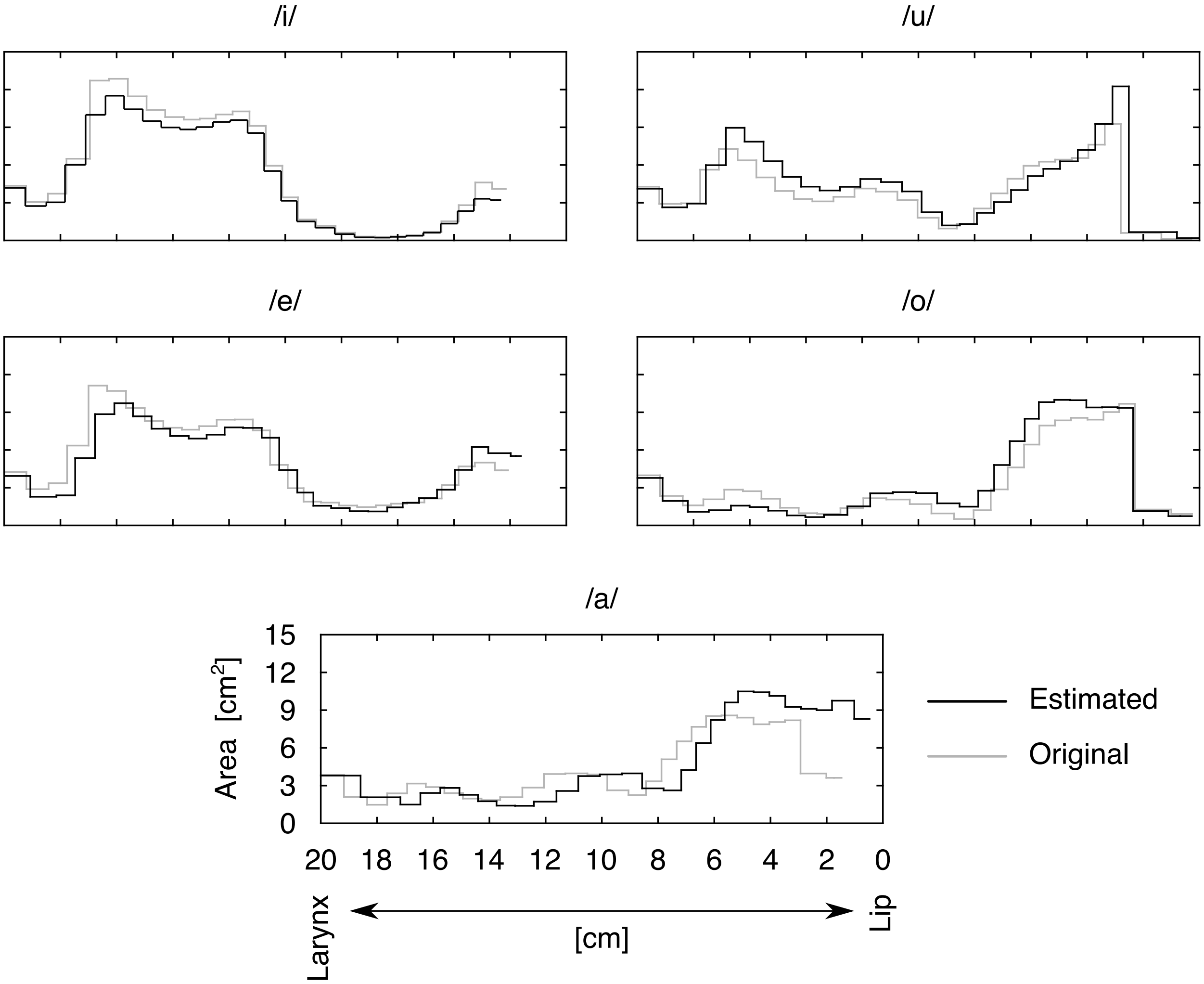

Figure 2.3: The area functions of the vocal tract for five vowels. The area functions derived from [http://www.cns.bu.edu/ speech/VTCalcs.php] and the inversely estimated ones are shown for each vowel.

2.3.2 Development of articulation

Developmental changes in articulatory parameters

To investigate changes in articulatory parameters over development, I calculated the mean values of the articulatory parameters and obtained developmental trends in these values

Table 2.2: The mean values of the articulatory parameters for each vowel at 12 months and the inversely estimated mean values of the articulatory parameters. For detailed procedure to obtain these values, see Sec. 2.2.5.

Vowel		Jaw	Tongue position	Tongue shape	Tongue apex	Lip aperture	Lip protrusion	Larynx height
/i/	original	-1.067	1.023	0.471	2.055	-0.006	-0.035	-1.209
	estimated	-1.056	1.304	0.173	1.910	1.236	-0.456	-1.091
/e/	original	-0.888	1.230	0.630	1.882	0.480	-0.206	-1.435
	estimated	-1.545	1.328	0.749	1.818	1.233	0.053	-0.978
/a/	original	-1.277	1.925	0.753	1.606	0.404	-0.148	-0.623
	estimated	-1.254	2.025	0.248	1.630	1.276	-0.282	-0.556
/u/	original	-0.917	1.587	0.135	1.489	-0.697	0.299	-1.238
	estimated	-0.406	1.647	-0.089	1.822	0.368	0.343	-1.335
/o/	original	-1.013	1.781	-0.171	1.334	-0.257	-0.122	-1.047
	estimated	-1.112	1.782	0.203	1.590	0.960	-0.205	-0.890

using the third-order polynomial function. As Fig. 2.4 shows, the relative order of values of the articulatory parameters is consistent with our standard. For instance, regarding the tongue position parameter, the polynomial function for each vowel was arranged in ascending order as follows: /i/, /e/, /u/ and /o, a/. This order, which is high-front (/i/), mid-front (/e/), high-back (/u/), and mid and low vowel (/o, a/), met our standard for tongue position movements (see also Fig. 2.1A). As for the tongue apex parameter, the functions for /i/ and /e/, which are articulated at the anterior of the vocal tract, were larger than the others. The arrangement of the lip aperture parameter also meets our standard: /u/ and /o/ are articulated with a narrow lip aperture. Altogether, the mean values of the articulatory parameters tended initially to be biased toward positive or negative values, and became closer to zero over development.

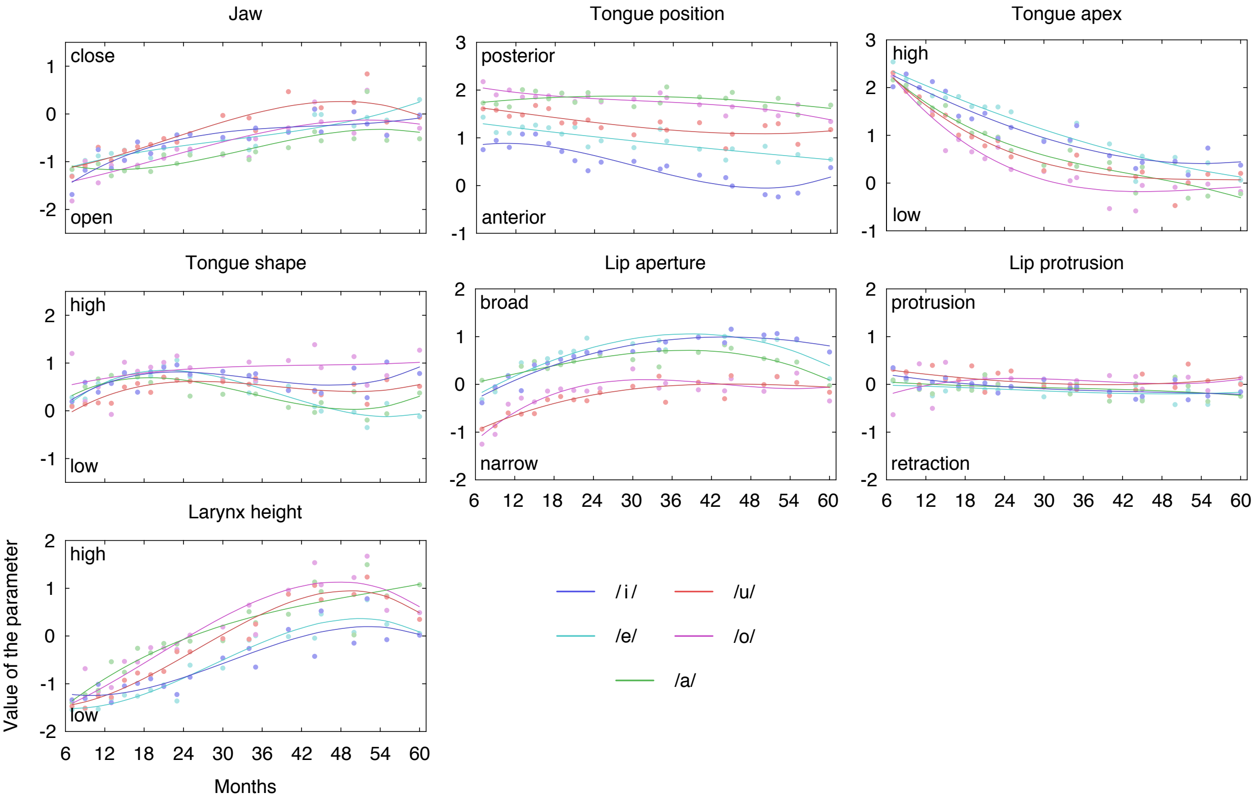

Figure 2.4: Mean values of the articulatory parameters as a function of months. The curve lines were obtained by the third-order polynomial function with the months of age as the independent variable and mean values as the predictors.

Analysis of distinctness

I investigated the distinctness of Japanese vowel clusters in children's speech. In order to evaluate distinctness, I calculated the correct classification rate obtained from LDA. Figure 2.5 shows results of LDA for children's speech in the articulatory space. As part of LDA, I conducted stepwise selection for the independent variables (p <0.05) and recruited significant variables into the classification function. Based on these recruited independent variables and the correct classification rates, the developmental course from 6 to 60 months was divided into three stages. Until 9 months of age, Japanese vowel clusters were mainly determined by the tongue position and lip aperture parameters. Then, until 17 months of age, the clusters became more distinct. This differentiation was caused by the recruitment of the jaw and tongue apex parameters. Subsequently, the recruitment of the tongue shape parameter resulted in more distinct clusters. The correct classification rates improved in increments of about 10%. This means that, as the stages proceeded, children tended to produce each vowel in a more distinct articulatory states. I also analyzed distinctness in the acoustic space and show the result in Fig. 2.6. As the children grew up, the correct classifi-

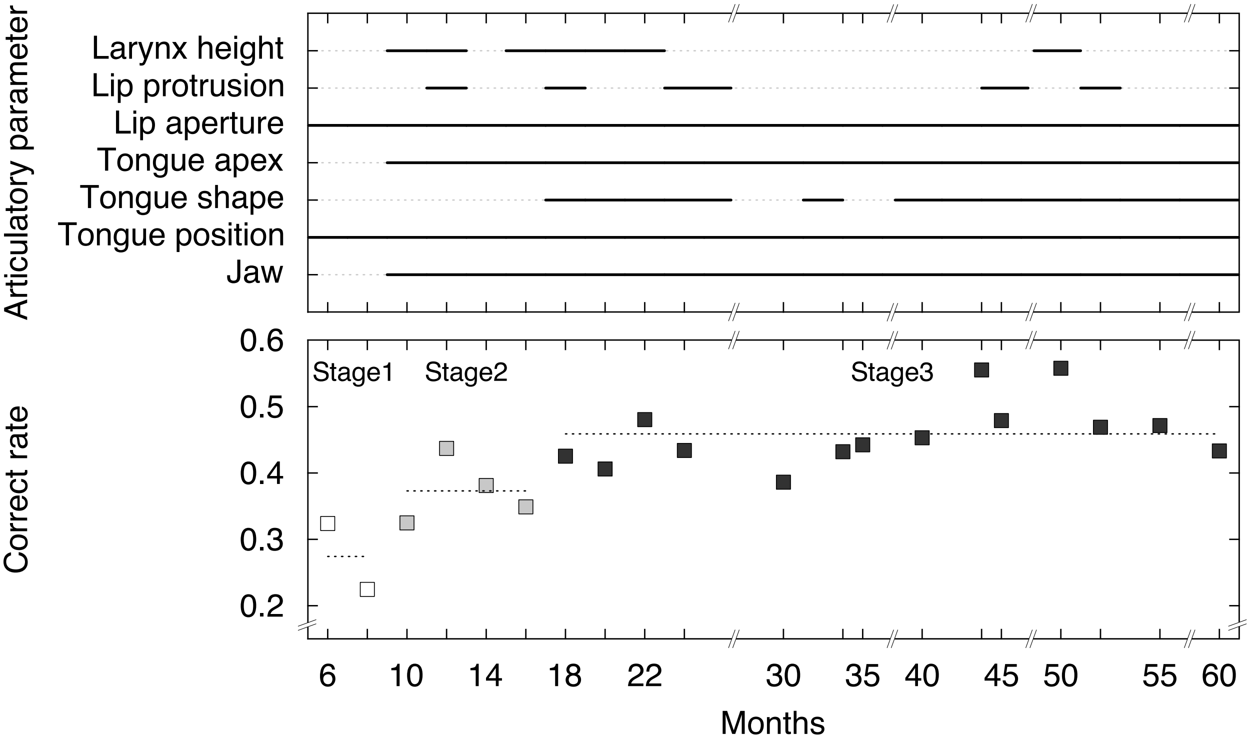

Figure 2.5: The distinctness of Japanese vowel clusters evaluated by LDA in the articulatory space. The upper panel indicates significant parameters for vowel classification. Solid lines in the upper panel mean that these parameters significantly contribute to classification (p <0.05). Correct rate was plotted as a function of age (in months). Dashed lines in the lower panel show averages of the correct classification rates at each stage.

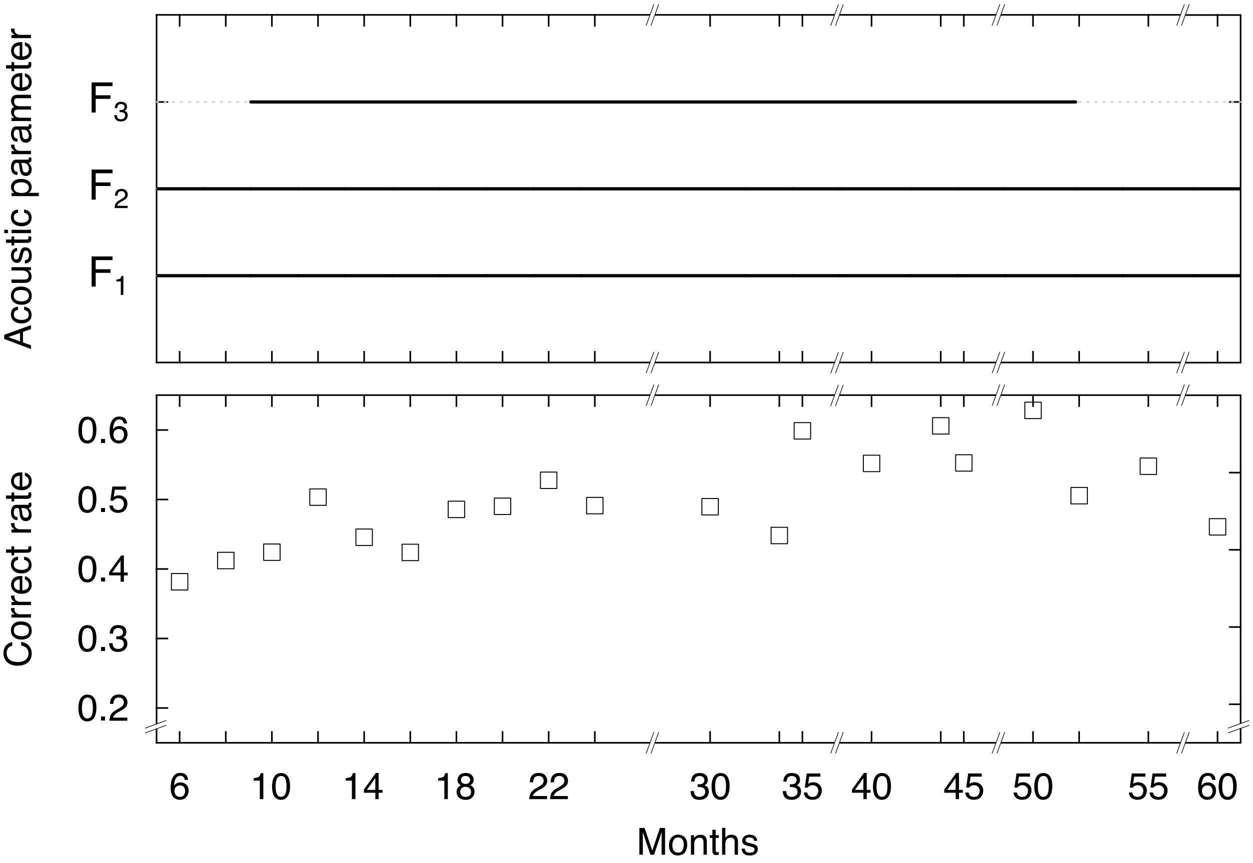

Figure 2.6: The distinctness of Japanese vowel clusters in the acoustic space as evaluated by LDA. The upper panel indicates significant parameters for vowel classification. Solid lines in the upper panel mean that these parameters significantly contribute to the classification (p <0.05). The lower panel indicates the correct classification rate as a function of age (in months).

cation rate improved. Together with the result of LDA in the articulatory space (Fig. 2.5), it can be seen that as the rates in the articulatory space improved, the rates in the acoustical space increased. Incidentally, the mean absolute difference between the correct classification rates in the articulatory space and those in the acoustical space was 0.076 (1 S.D. is ±0.040).

To examine whether the recruitment of articulatory parameters observed in the group analysis were common developmental patterns among children or not, I further conducted

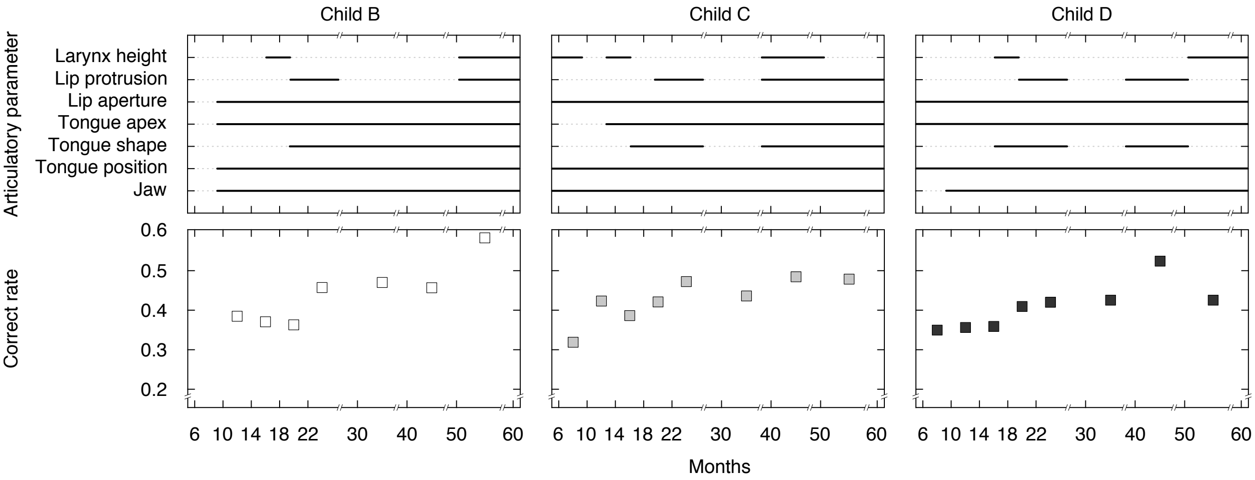

Figure 2.7: Individual analysis of the dinstinctness of Japanese vowel clusters evaluated by LDA in the articulatory space. The upper panel indicates significant parameters for vowel classification. Solid lines in the upper panel mean that parameters significantly contribute to classification (p <0.05). Correct rate was plotted as a function of age (in months). Note that, because of lack of samples, I could not perform LDA for child B at 6 months.

the individual analysis of distinctness in the articulatory space. As shown in Fig. 2.7, for all three children, two common developmental patterns were observed. First, the tongue position and lip aperture parameters both contributed significantly to the different vowel productions from early development. Second, the tongue shape parameter contributed to the vowel productions only from relatively later development. In addition, there were some differences among the three children: Although child B showed early contribution of both the jaw and tongue apex parameters, child C and D showed early contribution of one of those parameters. Overall, these individual developments support the division into three stages observed in the group analysis.

For detailed results of LDA at each stage, Fig. 2.8 shows the confusion matrices for LDA (upper panels) and the logarithmic values of likelihood ratio to be classified into the correct vowel category over the others (lower panels) at each of the three developmental stages. The values of the diagonal cells of the confusion matrix are as follows: /i/: (stage 1, stage 2, stage 3) = (0.429, 0.399, 0.548), /e/: (0.072, 0.301, 0.353), /a/: (0.353, 0.380, 0.427), /u/: (0.261, 0.421, 0.413) and /o/: (0.714, 0.417, 0.491). At the first stage (6–9 months old), articulatory states were divided into the three clusters: central (/a/), front (/i, e/) and back (/u, o/) vowels. Subsequently, at the second stage (10–17 months old), these three clusters were differentiated into five by means of a back-vowel cluster split into /o/ and /u/ clusters and a front-vowel cluster split into /i/ and /e/ clusters. Eventually, at the third stage (18

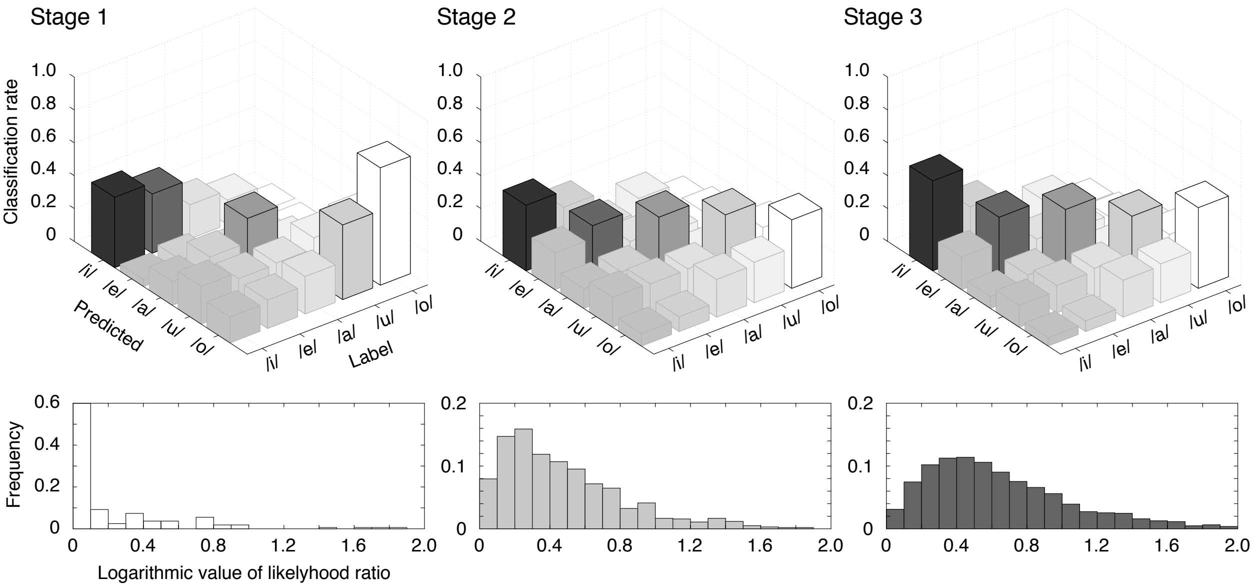

Figure 2.8: The confusion matrices for LDA (upper panels) and distributions of logarithmic values of likelihood ratio (lower panels) at the three developmental stages. In the upper panel, the cells with maximal classification rates for each label are highlighted by thick lines. In the lower panel, larger logarithmic values for the likelihood ratio mean that the probability of being classified into a correct vowel cluster is higher and that of being classified into the wrong cluster, lower. For details of the calculations for the logarithmic values of the likelihood ratio, see Sec. 2.2.7.

Table 2.3: The values of the indices quantifying the properties of clusters. The clusters were estimated by variational Bayesian inference (VB) with three independent components. The indices aim to quantify the degree to which clusters occupied the whole of all articulatory states (share of clusters), whether each cluster is consistent with a single vowel category (purity within a cluster), and whether a set of clusters explains all the vowel categories (completeness of vowels). For detailed definition of each index, see Sec. 2.2.7.

Stage	Share	Purity	Completeness
	of clusters	within a cluster	of vowels
1	0.805	2.438	0.700
2	0.970	1.553	0.913
3	0.966	1.985	0.960

months old and older), each of these clusters was refined and became more distinct. In order to evaluate these differentiation and refinement process statistically, I conducted Welch's t-test for logarithmic values of likelihood ratio and adjusted p-values by multiplying them by three (${}_3C_2 = 3$). The results showed that the sharpness of the clusters was significantly different among the three stages and became larger with development ($p < 0.01$).

Analysis without vowel labels attached by transcribers support these developmental patterns. In the present study, we used the three indices to quantify the properties of the clusters: share of clusters, purity within a cluster and completeness of vowels, which assessed how much all clusters occupied the overall articulatory states, whether each cluster is consistent with a single vowel category, and whether a set of clusters explains all of a vowel categories, respectively. Table 2.3 shows the three indices at the three identified stages. The values for share of clusters and completeness of vowels increased between stages 1 and 2. These results support the presence of the differentiation process implied by the LDA. In contrast with these two indices, purity within a cluster decreased between stages 1 and 2 and increased between stages 2 and 3. This implies that samples in a cluster tend to be more similar to each other at stages 1 and 3 than at stage 2.

2.4 DISCUSSION

Above, I have described developmental changes in articulatory state during vowel production on the basis of the acoustic-to-articulatory inversion technique.

As shown in the longitudinal changes in the mean values of the articulatory parameters, the distribution of the articulatory parameters was biased toward positive or negative values in early development and became closer to zero with ages. These biased distributions would disagree with the assumption of the previous study [Serkhane *et al.* 2007] that infants can adopt a full range of parameter values for the tongue position, tongue shape, jaw and lip height. Moreover, since the jaw parameters tended to be biased toward negative values, the jaw contributed to opening of the oral cavity. This means that the jaw works vowel production in the early development of speech production during the babbling stage.

The group analysis revealed that articulatory development goes through three developmental stages, in which early forms of vowels are differentiated into the vowels of adult Japanese. Subsequent individual analysis of distinctness among vowels showed results consistent with the developmental patterns observed in the group analysis, in which three children shared patterns of parameter recruitmemnt: early contribution of the tongue position and lip aperture parameters to different vowel production, and later contribution of the tongue shape parameter. These results of the individual analysis potentially support the division of developmental stages observed in the group analysis.

At the first stage (6–9 months), the tongue position and lip aperture parameters contribute significantly to different vowel productions. The confusion matrix for LDA suggests that articulatory states at this stage are divided into three vowel clusters, that is, central (/a/), front (/i, e/) and back (/u, o/). These three clusters correspond to the corners of the F_1–F_2 vowel space. This result implies that, in early development, articulatory states do not adjust to Japanese vowel structure but reflect mere coordination between tongue and lip articulations. Although kinematic studies [Green *et al.* 2002, Nip *et al.* 2009] measuring jaw and lip motions shows that control of the jaw matures earlier than that of the lip, the present study suggests that regulation of lip aperture is nevertheless already functioning at this point to generate different vowels. As we described above, Fig. 2.4 implied that the jaw worked vowel production in the early development of speech production during the babbling stage. Yet, the jaw movements at this stage would be a mere open-close alternations. In other words, the jaw movements at this stage would not be controlled to a different degree depending on kinds of vowels. In the mandibular oscillation theory [MacNeilage & Davis 2000, MacNeialge 2008], three preferred CV sequences (labial–central, coronal–front and dorsal–back sequences) in infancy are assumed to result from the position of the tongue and the inertia caused by the close-open alternation of the jaw. In this sense, which of the three CV sequences is generated depends essentially on tongue position. Therefore, the jaw parameter does not contribute to the generation of different articulatory states from vowel to vowel.

In fact, at the following stage (10–17 months), which is the stage at which the first word emerges, the jaw parameter plays a significant role in the production of different vowels. At this stage, the three clusters of the first stage have differentiated into five ambiguous clusters. The recruitment of the jaw and tongue apex parameters contributes to this differentiation process, from a relatively universal vowel structure into the vowel structure of adult Japanese. As reported by one acoustical study [de Boysson-Bardies *et al.* 1989], infants' vocalizations begin to adapt to the phonetic properties of their native languages toward the end of the first year of life. This should be reflected in the stage-like changes identified in the involvement of articulatory parameters by this study. At stage 2, in addition to the jaw parameter, the tongue apex parameter is recruited into the classification function. Tongue apex position is assumed to account for the apical gesture during dental consonants and high-front vowels [Maeda 1990]. Together with the confusion matrices at stages 1 and 2, the recruitment of the tongue apex parameter would contributes to the differentiation of front vowels into /i/ and /e/. In addition, at the third stage (18 months and older), the recruitment of the tongue shape parameter leads to more distinct vowel productions than previously. As a result, the ambiguous clusters are refined into distinct ones. The presence of these differentiation and refinement process is supported not only by the correct classification rates yielded for LDA but also by the distributions of likelihood ratios to being classified into the correct vowel category over the others (lower panels in Fig. 2.8).

According to a previous study [Takano & Honda 2007], tongue movements for vowel production are governed by three extrinsic lingual muscles, that is, the genioglossus, hyoglossus, and styloglossus. The genioglossus and styloglossus are divided into anterior, middle, and posterior parts, and the hyoglossus into anterior and posterior parts only. Articulatory movements controlled by the tongue position parameter are governed by the posterior genioglossus and hyoglossus, while movements controlled by the tongue shape parameter are governed by the anterior genioglossus and middle styloglossus [Honda 1996]. Thus, the result showing earlier significance of the tongue position parameter and later significance of the tongue shape parameter might reflect a difference in maturational speed between the respective muscles in vowel production. That is, synergies between the posterior genioglossus, which has the largest volume and can apply the greatest force [Takemoto 2001], and the hyoglossus may mature earlier than those between the anterior genioglossus and styloglossus.

When it comes to anatomical changes in the vocal tract, the later recruitment of the tongue shape parameter would be related to laryngeal descent [Fitch & Giedd 1999,Goldstein 1980,Sasaki *et al.* 1977,Vorperian *et al.* 1999,Vorperian *et al.* 2005]. The narrow pharyngeal cavity in early development could prevent the tongue movements necessary to change the tongue shape parameter; in this sense, laryngeal descent would affect pharyngeal cavity size and therefore enable children to change the tongue shape parameter. In fact, a previous study [Vorperian *et al.* 2005] reports a breakpoint in laryngeal descent at around 18 months, which coincides with the beginning of stage 3 as defined in the present study.

The unsupervised classification of inversely estimated articulatory parameters supports the plausibility of above differentiation process. To quantify the result of the classification, I calculated three indices—share of clusters, purity within a cluster, and completeness of vowels—and found that share of clusters and completeness of vowels increased from stages 1 to 2, while purity within a cluster decreased from stages 1 to 2 and then increased from stages 2 to 3. Taken altogether, this means that at the first stage, articulatory states can be divided into fewer clusters, which represent only a part of Japanese vowels. Subsequently, at the second stage, the number of clusters increased and clusters corresponding to more vowels appeared. Yet each cluster was interwoven with multiple vowel categories. Eventually, at the third stage, each cluster came to be constructed primarily by a distinct vowel specific to Japanese.

One simulation study [Liljencrantz & Lindblom 1972] proposes the adaptive dispersion principle of constructions of the vowel system: the distinctive sounds of language tend to be positioned in the phonetic space so as to maximize perceptual contrast. The present study further shows that although the distinctness of vowels as calculated from articulatory parameters was slightly lower than that calculated from acoustic parameters, articulatory dispersion is consistent with perceptual dispersion. For instance, the tongue position and lip aperture parameters construct three vowels, which are consistent with the corners of the perceptual vowel space.

Although this study focuses only on static articulatory states during a single vowel production, the combination of units such as phonemes, syllables and words is crucial for language. One simulation study [Zuidema & deBoer 2009] proposes that the combinatorial properties can be determined based on a maximization of acoustical distinctness. Yet, in reality, as shown in the preference of CV patterns in early development [MacNeilage & Davis 2000,MacNeialge 2008], articlatory factors constrain the combination of these linguistic units. Similarly, not only the anatomical structures of the vocal tract but also articulatory control system constraints the exploration process, that enables us to obtain a forward model to map the articulatory space to the acoustic space. In other words, speech production should stabilize in accord with a potential function that has perceptual dispersion and coordination among articulators as variables. Aside from these within-individual factors, communicative interactions with others also affect the emergence of the structure of language [Kirby *et al.* 2008, Nowak & Krakauer 1999].

Two previous studies criticize the application of the adult articulatory model to young children [de Boer & Fitch 2010, Lieberman 2012]. They point out the midsagittal mismatch of vocal tract shape between adults and young children, which may be caused by difference of slope in the vocal tract. Another criticism is that linear scaling of the vocal tract can cause unrealistic deformation of the space. However, I do not claim that each articulatory parameter runs through the full range of the codomain. In fact, the distribution of each parameter is biased toward some area. Nevertheless, it seems clear that in order to acquire a more precise understanding of articulatory development, further studies on articulatory dynamics during speech production using methods such as ultrasound are needed.

2.5 CONCLUSION

I described the application of an acoustic-to-articulatory inversion technique to identify and analyze the development of vowel articulation. First, I validated inversely estimated articulatory parameters. Although the classical study in this area proposed that infants start by vocalizing all possible speech sounds of the world's languages [Jakobson 1968], other studies have shown that infants produce only limited kinds of speech sounds [MacNeilage & Davis 2000, MacNeialge 2008, Oller 2000], suggesting that their speech production is constrained by the developing anatomy of the articulatory organs and by the rate of maturation of motor control. I attempted to solve these issues by estimating articulatory states using a scalable model and conducting acoustic-to-articulatory inversion without any assumption regarding the distribution of any parameter. I conducted evaluations of the coordination of articulatory parameters of vowel production using LDA and VB. Both evaluations suggested that development of articulatory coordination for vowel production goes through three stages as recruitment of the different articulatory parameters causes changes in articulatory coordination and in the resultant vowel sounds. Aspects of the inherent nature of articulatory coordination will determine the course of the early development of speech production. The present study suggests that the initial vowel articulations are founded on coordination between lip aperture and tongue position, and that maturation of tongue muscle coordination and vocal tract anatomy are critical in the differentiation and refinement, over the course of development, of vowels that are properly adjusted to the native language.

These results show the potential of my approach to languages other than Japanese. Although the present study focused only on static articulatory states, the extension of the approach to understand the dynamics of the development of articulatory coordination during speech production also holds promise.

Chapter 3

Development of a Serial Order in Speech Constrained by Articultory Coordination

3.1 INTRODUCTION

Speech production is very complex; the central nervous system coordinates more than 100 muscles to control movements of jaw and soft tissues such as the lips and tongue [Golfinopoulos *et al.* 2010, Kent 2004]. This property enables us to perform an extraordinarily rapid sequence of action, that is, the production of dozens of syllables or phonemes in a single breath. This phenomenon is related to the serial order problem [Lashley 1951], which addresses how behaviors are sequenced without triggers from sensory feedback. The serial order in speech production has been a central issue in the phonological study that attempts to find universal structures that govern all human languages [Chomsky & Halle 1968]. Other lines of study focusing on motor control of articulatory systems have argued that discrete phonemic units in human languages are founded on the dynamical nature of the neuromuscular system that controls precise coordination among many articulators [Browman & Goldstein 1990, Fujimura 1981, Kelso *et al.* 1984, Rochet-Capellan *et al.* 2007, Trembley *et al.* 2003]. Moreover, studies on the acquisition of speech production have attempted to interpret phonological structures preferred in language and babbling in terms of the articulatory system [Locke 1983, MacNeilage & Davis 2000, MacNeialge 2008, Nam *et al.* 2009]. Aside from neuromuscular coordination, vocal tract anatomy in early development differs from those of adults in that infants' vocal tracts are not only smaller than adults', but they have a broader oral cavity, a tongue mass that is proportionally larger and more anterior, and a more gradually sloping pharyngeal tract [Vorperian *et al.* 2005]. Findings from these studies suggest that a neuromuscular coordination and the vocal tract anatomy constrain the development of phonotactics for words in a language.

The human articulatory system consists of organs, such as the jaw, tongue and lips (Fig. 3.1A). In order to produce speech sounds, speakers coordinate these organs to change the vocal tract configuration. In brief, these articulatory movements are close-open alternations of the vocal tract that generate series of consonant and vowels.

Vowels are mainly characterized by the horizontal and vertical positions of the tongue. According to the horizontal positions of the tongue, vowels are classified into front (*e.g.*, [i] end [e]), center (*e.g.*, [a]) and back vowels (*e.g.*, [u] and [o]). According to the vertical positions of the tongue, vowels are classified into high (*e.g.*, [i] and [u]), middle (*e.g.*, [e] and

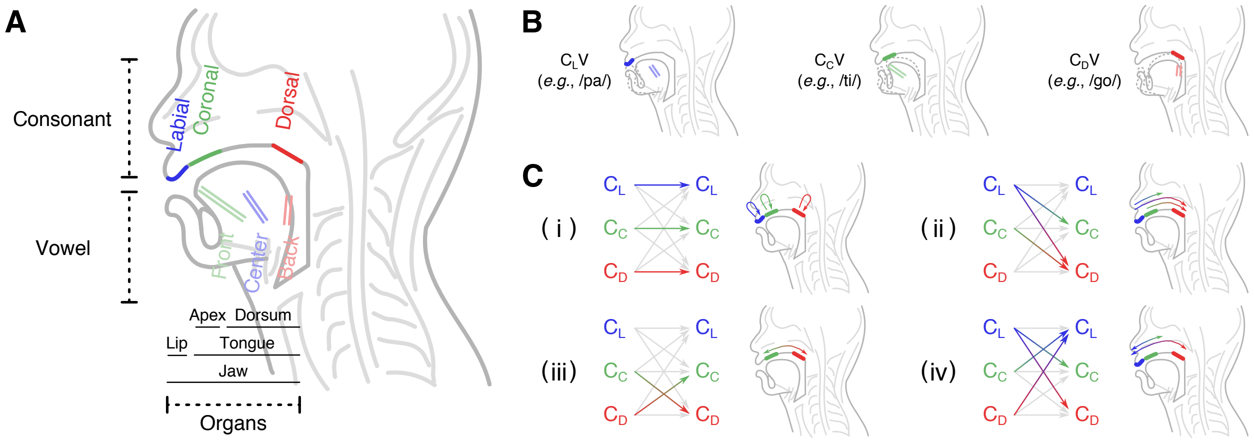

Figure 3.1: The classification of serial order in articulations in the previous and present studies. (A) The place of articulations for consonants and vowels, and articulatory organs involved in each consonant. Depending on the horizontal position of the tongue, vowels are categorized into three types including front, center and back. This figure illustrates three places of articulations, including labial, coronal and dorsal. Labial consonants are mainly articulated by the lips and jaw. Coronal consonants are mainly articulated by the tongue apex and jaw. Dorsal consonants are mainly articulated by the tongue dorsum and jaw. (B) Three consonant-vowel patterns preferred by infants in early development. Focusing on three consonantal and vowel categories, theoretically speaking, it is possible to produce nine consonant-vowel sequences. However, infants prefer three out of those nine possible sequences: labial-center, coronal-front, and dorsal-back [MacNeilage & Davis 2000,MacNeialge 2008]. (C) Serial order in articulation of consonants in consonant-vowel-consonant(-vowel) sequences. In the present study, focusing on the relationship among articulators producing adjacent consonants, I divided sequences into four categories: (i) Sequences consists of consonants produced at the same place of articulation. (ii) Sequences produced by movements from more anterior place to posterior one. (iii) Sequences consist of coronal and dorsal consonants, which are articulated by the same organ but different places (intra-organ articulations). (iv) Sequences consist of labial and coronal/dorsal consonants, which are articulated by different organs: lips and tongue (inter-organ articulations).

[o]) and lower vowels (*e.g.*, [a]). Consonants can be classified into categories according to their places of articulation. Of these, labial (*e.g.*, [p] and [m]), coronal (*e.g.*, [t] and [d]) and

dorsal consonants (*e.g.*, [k] and [g]) are the major categories. Consonants are also classified by the manner of articulation, such as fricative (*e.g.*, [s] and [z]), nasal (*e.g.*, [m] and [n]), and stop (*e.g.*, [p] and [t]). Consonants are produced as a result of the coordinated movements of one or more articulators; these movements are referred to as consonantal gestures. The lips, tongue apex and tongue dorsum articulate to produce labial, coronal and dorsal consonants, and the jaw can contribute to all of these consonants. Both coronal and dorsal consonants are produced by the tongue but they are governed by different tongue tissues. Regarding the development of speech sounds, the previous study [MacNeialge 2008] has shown that only a small set of sounds are used frequently in babbling. Stop and nasal consonants with labial, coronal and dorsal places of articulation and lower-front and lower-center vowels are present in infants from all language environments from the beginning of the babbling stage and remain in place in the first-word stage.

As children acquire speech production, previous analyses have revealed a shared preference for specific articulatory patterns [MacNeilage & Davis 2000, MacNeialge 2008]. For example, specific consonant-vowel (consonants and vowels are hereafter described as C and V, respectively) patterns are observed at significantly high frequencies in the period of babbling (Fig. 3.1B). Regarding the order of consonant in a CVC(V) sequence, the same consonantal gestures are repeated (Fig. 3.1C-i). Studies have also shown that both children and adults produce labial-vowel-coronal sequences more frequently than coronal-vowel-labial [MacNeilage & Davis 2000, MacNeialge 2008]. A previous study [Ingram 1974] showed more generally that the first consonant in words produced by children aged between 16 and 24 months has a more anterior place of articulation than the second one (Fig. 3.1C-ii), and this pattern has been referred to as fronting. These patterns in child speech production are observed not only in many modern languages but also in proto-language corpora [MacNeilage & Davis 2000, MacNeialge 2008]. Note that there is no preference for the sequences in adult Japanese [MacNeilage *et al.* 1999, Tsuji *et al.* 2012].

For speech production, one articulator must coordinate with other articulators and this coordination can vary from phoneme to phoneme. Therefore, not only simultaneous coordination of articulatory organs but also their sequential coordination is critical for a serial order in speech production. Focusing on neuromuscular coordination, previous studies emphasize the important roles of the jaw in development of speech production [MacNeilage & Davis 2000, MacNeialge 2008]. The preferred CV patterns and repetitions of CVC(V) sequences in babbling are generated by rhythmic jaw cycles on which speech production is founded. According to this explanation, a resting position of the tongue and an inertia caused by close-open alternation of the jaw generate the CV patterns, such as /papa/, /dede/ and /gogo/. Among the CV patterns preferred in babbling, these studies [MacNeilage & Davis 2000, MacNeialge 2008] especially focus on sequences consisting of labial consonants and center vowels supposing that mere jaw raising gestures would induce a labial closure of the vocal tract. Other studies [Vilain *et al.* 1999] also show that the mandibular oscillations with passive tongue configurations could induce a coronal closure of the vocal tract. Furthermore it has been speculated that the mandibular oscillations can be controlled by the central pattern generators, which may work in the brain stem of infants from early on [Gracco & Abbs 1988, Grillner 1982, MacNeialge 2008, Netsell 1982, Wilson *et al.* 2008]. As development proceeds, independent and active recruitment of the different part of the tongue and the lips into the cycle causes increases in the variety of lingual consonants used in speech production. Regarding this divergent process, the previous studies mainly focus on the ordering of labial and coronal consonants [MacNeilage & Davis 2000, MacNeialge 2008] (Fig. 3.1C-ii), which is produced on the basis of the inter-articulatory coordination between the lips and tongue apex. Previous kinematic studies [Goffman & Smith 1999, Green *et al.* 2000, Smith & Zelaznik 2004] also show that development of the coordinative organization of the jaw and lips during speech production for a specific sound can be characterized by distinct phases, in which movements of the jaw become differentiated and integrated with movements of the lips.

In addition to the repetition and fronting that dominate early speech, the present study examines the longitudinal development of CVC(V) sequences produced by serial coordination of articulators such as the lip, tongue apex and tongue dorsum. In particular, I focus on an intra-articulator and an inter-articulator serial order (Figs. 3.1C-iii and iv). An intra-articulator serial coordination between the tongue apex and dorsum (Fig. 3.1C-iii) might be constrained by a physical connection within the tongue tissue and/or by a limited neural control of articulators in the period of babbling, and are established through the later development of the neuromuscular system. An inter-articulator serial coordination between the lips and tongue dorsum as well as between the lips and tongue apex (Fig. 3.1C-iv) might be also constrained by a developing mechanism for the neuromuscular control of these articulators.

If the intrinsic property of the neuromuscular system constrains early developments of speech production, language universal developmental patterns could be observed. On the other hand, lexical specifications and linguistic inputs facilitate infants' speech production of their native language [de Boysson-Bardies 1999, Fikkert & Levelt 2008]. In order to distinguish between neuromuscular maturation and linguistic stimulations on development of speech production, cross-linguistic comparison between different languages and analysis of child-directed speech are required.

To that end, I used a Japanese corpus, which aimed to longitudinally track early normal speech development; children and their parents were not required to do any particular task other than to record their talk during everyday situations in daily life [Amano *et al.* 2009]. The database has been used to reveal longitudinal changes in fundamental frequency, formant frequencies and conversational style [Amano *et al.* 2006, Ishizuka *et al.* 2007, Kajikawa *et al.* 2004, Kajikawa *et al.* 2004]. Thus, I further studied CVC(V) sequences reflecting the coordination of articulators. In the same way, I performed analysis using an English database [Davis 2007], which is provided in the CHILDES projects [MacWhinney 2000]. This database also aimed to record infants' naturally occurring vocalizations in their home environment, usually in the company of a parent and/or siblings [Matyear *et al.* 1998]. Through data collections for this database, the previous studies have revealed that some CV patterns and labial-coronal fronting patterns are preferred in early development [MacNeilage & Davis 2000, MacNeialge 2008]. In the present study, I investigate longitudinal changes in CVC(V) productions in Japanese and English to examine whether serial coordination of articulators acts as a constraint for the developmental emergence of sequences in words.

3.2 METHODS

Data for this study are taken from the NTT Japanese infant speech database [Amano *et al.* 2009] and the Davis corpus for English [Davis 2007, MacWhinney 2000]. The Japanese corpus longitudinally tracks five children from birth to 60 months old. The English corpus contains data of 21 children and tracking periods of it ranges from five to 27 months. For both databases, well-trained transcribers listened to the speech files and transcribed utterances, and checked reliability (for details see [Amano *et al.* 2009, Matyear *et al.* 1998]). The utterances were segmented from continuous speech based on the pause duration. In the present study, I used transcriptions of utterances by three children aged 7 to 60 months (Table 3.1), transcriptions of child-directed speech by their parents (Table 3.2), and duration data for each phoneme for Japanese. Because of lacks of the transcriptions, I did not use data of the other two children contained in the Japanese corpus. I also used transcription of utterances by 21 children aged 7 to 37 months for English (Table 3.3). From the transcriptions, I extracted CVC units contained in sequences such as CVC and CVCV, and examined changes in place of articulation patterns through stages of development. For instance, in the case of $C_1VC_2VC_3$, I extracted two sequences C_1VC_2 and C_2VC_3. In this study, I focused on only three types of consonants: labial stop and nasal consonants ([p], [b] and [m], hereafter C_L), coronal stops and nasals ([t], [d] and [n], hereafter C_C) and dorsal stops ([k] and [g], hereafter C_D), which are favored from the beginning of the babbling stage [MacNeialge 2008]. Therefore, CVC sequences were classified into nine categories; C_LVC_L, C_LVC_C, C_LVC_D, C_CVC_L, C_CVC_C, C_CVC_D, C_DVC_L, C_DVC_C and C_DVC_D. Consonants and vowels have different characteristics: consonant production has a discrete character generated by transient coordination of articulators and vowels have a relatively continuous character primarily controlled by position of the tongue dorsum. Given this difference, I assumed that the CC relationship in CVC(V) sequences may primarily exhibit combinatory complexity as a source of linguistic complexity. Therefore, I omitted vowels from the present analysis and focused only on the

CC sequences. During the extraction process, syllable positions in an utterance were not taken into account, and syllable boundaries and diacritic marks were ignored. All CVC sequences were analyzed to determine the token frequencies of serial order relationship.

3.2.1 Analysis of individual data for Japanese

For the Japanese analysis, taking into account that the corpus tracks a small number of subjects over an extensively longitudinal period [Amano *et al.* 2009], I first conducted individual analysis. I calculated ratios of each occurrence of the nine CVC categories to the total CVC occurrences for each child for each month of age. To analyze serial order in places of articulation for consonants, I calculated four types of developmental curves (Fig. 3.1C): (i) repetitive articulation at the same place of an organ ($\mathrm{C_LVC_L} + \mathrm{C_CVC_C} + \mathrm{C_DVC_D}$), (ii) articulatory movements from more anterior places ($\mathrm{C_LVC_C} + \mathrm{C_LVC_D} + \mathrm{C_CVC_D}$) to posterior ones ($\mathrm{C_CVC_L} + \mathrm{C_DVC_L} + \mathrm{C_DVC_C}$), (iii) different places of articulations within the same organ ($\mathrm{C_CVC_D} + \mathrm{C_DVC_C}$) and (iv) articulations by different organs (Figs. 3.1C-i to -iv). Traditionally, linguistics treat the tongue apex and the tongue dorsum differently (the tongue apex and dorsum are used to articulate coronal consonants and dorsal consonants, respectively). Thus, in terms of (iv), I calculated two values, that is, $\mathrm{C_LVC_C} + \mathrm{C_CVC_L}$ and $\mathrm{C_LVC_D} + \mathrm{C_DVC_L}$. In the following, I referred to (i)–(iv) as the repetitions, fronting, intra- and inter-organ articulations.

For (i), I defined and normalized a developmental curve. In detail, I defined the developmental curves by using kernel regression (Alg. 2) [Bishop 2007]. Although there are many options of curve fitting such as polynomial function, I chose the kernel regression to obtain

Table 3.1: The number of CVCs in the Japanese corpus [Amano *et al.* 2009]. Each number indicates the number of occurrences of each type of CVC (Group: number of samples in the group data, Individual: number of samples in the individual children, Total: number of samples of total CVC(V) sequences).

Month	Repetitions				Intra-organ				Inter-organ								Total
									Labial-Coronal				Labial-Dorsal				
	Group	Individual			Group	Individual			Group	Individual			Group	Individual			
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	8	8	0	0	1	1	0	0	0	0	0	0	0	0	0	0	9
10	4	0	4	0	2	0	2	0	0	0	0	0	1	0	1	0	7
11	4	0	1	3	2	0	2	0	0	0	0	0	0	0	0	0	6
12	10	5	4	1	2	0	2	0	0	0	0	0	0	0	0	0	12
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
14	11	9	1	1	4	4	0	0	5	4	0	1	5	5	0	0	25
15	12	7	0	5	1	1	0	0	0	0	0	0	0	0	0	0	13
16	40	30	2	8	4	2	2	0	0	0	0	0	0	0	0	0	44
17	51	17	1	33	2	2	0	0	1	0	1	0	2	2	0	0	56
18	43	39	4	-	22	21	1	-	11	10	1	-	3	3	0	-	79
19	60	39	4	17	7	4	2	1	4	3	0	1	5	4	0	1	76
20	72	59	13	-	18	10	8	-	10	5	5	-	10	8	2	-	110
21	61	35	-	26	10	10	-	0	8	7	-	1	8	8	-	0	87
22	55	46	9	-	19	9	10	-	16	10	6	-	31	27	4	-	121
24	182	49	109	24	190	33	141	16	132	26	94	12	93	19	44	30	597
25	89	42	47	-	86	42	44	-	55	35	20	-	46	29	17	-	276
30	80	23	41	16	70	21	48	1	42	20	22	0	33	18	14	1	225
34	51	-	47	4	58	-	53	5	16	-	16	0	14	-	12	2	139
35	78	60	18	-	92	84	8	-	23	20	3	-	18	16	2	-	211
40	155	87	17	51	211	118	28	65	89	57	8	24	52	32	2	18	507
44	19	-	19	-	35	-	35	-	22	-	22	-	4	-	4	-	80
45	260	179	3	78	347	239	7	101	133	92	2	39	68	42	0	26	808
50	28	28	-	-	50	50	-	-	32	32	-	-	8	8	-	-	118
52	147	-	55	92	149	-	78	71	83	-	25	58	58	-	4	54	437
55	53	53	-	-	57	57	-	-	25	25	-	-	6	6	-	-	141
60	84	84	-	-	77	77	-	-	64	64	-	-	11	11	-	-	236

higher determination coefficients (R^2) for fitting developmental trends. First, I calculated the Gramian matrix by using the radial basic function as the kernel function. The parameter

Table 3.2: The number of CVCs in child-directed speech in the Japanese corpus [Amano *et al.* 2009]. The notation is the same as that used in Table 3.1.

Month	Repetitions	Intra-organ	Inter-organ		Total
			Labial-Coronal	Labial-Dorsal	
7	66	40	7	5	118
8	35	33	20	3	91
9	73	39	55	7	174
10	22	54	35	10	121
11	43	45	52	9	149
12	141	151	159	30	481
13	85	149	82	27	343
14	87	137	60	29	313
15	152	146	92	23	413
16	153	142	144	30	469
17	217	231	120	46	614
18	231	280	164	39	714
19	164	143	108	26	441
20	226	206	98	26	556
21	108	160	59	22	349
22	218	181	146	62	607
24	222	220	107	36	585
25	46	43	33	9	131
30	48	39	34	15	136
34	66	57	16	5	144
35	124	123	40	13	300
40	10	15	2	1	28
44	3	6	0	1	10
45	64	49	34	9	156
52	15	26	11	5	57

Table 3.3: The number of CVCs in the English corpus [Davis 2007,MacWhinney 2000]. The notation is the same as that used in Table 3.1.

Month	Repetitions			Intra-organ			Inter-organ						Total
							Labial-Coronal			Labial-Dorsal			
	Group	Individual		Group	Individual		Group	Individual		Group	Individual		
7	10	-	8	2	-	2	2	-	1	0	-	0	14
8	251	19	119	19	5	11	6	1	0	3	0	3	279
9	131	4	96	35	0	31	6	0	1	5	0	4	177
10	1044	-	446	48	-	22	18	-	4	1	-	0	1111
11	701	52	202	28	3	5	39	3	8	23	3	3	791
12	598	51	162	33	3	3	29	2	16	17	4	1	677
13	840	39	321	59	0	6	30	2	7	14	0	5	943
14	1449	13	152	134	2	7	122	2	16	17	0	3	1722
15	1165	105	190	17	0	1	94	3	20	25	0	9	1301
16	1507	31	33	89	3	2	441	2	28	161	0	11	2198
17	1489	113	17	168	28	17	417	7	23	126	5	10	2200
18	1222	62	12	144	18	13	288	7	8	159	6	11	1813
19	1003	29	6	171	3	14	257	17	6	196	4	4	1627
20	300	26	-	49	4	-	41	6	-	53	1	-	443
21	283	17	12	60	16	11	47	4	5	33	6	2	423
22	232	27	12	36	7	7	77	8	19	33	6	4	378
23	442	54	14	131	7	7	75	5	4	57	13	5	705
24	352	41	-	40	15	-	91	12	-	31	1	-	514
25	224	45	21	51	8	7	73	13	3	53	1	3	401
26	128	27	11	23	5	2	31	2	4	18	0	0	200
27	90	47	9	42	28	6	24	8	3	18	7	0	174
28	65	-	4	21	-	3	20	-	2	27	-	4	133
29	35	9	4	35	22	0	13	6	1	7	4	2	90
30	162	134	24	14	4	6	11	2	8	6	2	3	193
31	70	-	-	16	-	-	6	-	-	8	-	-	100
32	40	4	3	18	3	1	23	0	3	16	0	3	97
33	32	14	-	11	4	-	24	4	-	7	0	-	74
34	68	39	7	30	15	5	21	4	8	7	5	0	126
35	20	7	4	17	3	3	25	7	11	4	2	0	66
36	33	-	20	3	-	1	29	-	22	5	-	0	70
37	34	-	-	5	-	-	2	-	-	0	-	-	41

of the kernel function was decided by the leave-one-out cross-validation method. Next, I set random values into hyper-parameters, deciding on a variance of the weight coefficients and the noise. Then, I calculated means and variances of the Gaussian distribution for the predictor variables and the weight coefficients. Subsequently, I updated the hyper-parameters. Finally, I estimated the optimal predictor variables by the iteration of these processes. In order to normalize an argument trajectory, I divided the trajectory by a maximal value. The goodness-of-fit of the kernel regression for the developmental curve was assessed by the R^2.

For (ii), I defined developmental curves as follows. When I focus on directions of articulations, the nine CVC categories are also classified into two additional categories: (a) direction of articulations from front to back ($\mathrm{C_LVC_C} + \mathrm{C_LVC_D} + \mathrm{C_CVC_D}$) and (b) direction of articulations from back to front ($\mathrm{C_CVC_L} + \mathrm{C_DVC_L} + \mathrm{C_DVC_C}$). I calculated $[(a) - (b)]/[(a) + (b)]$ as the fronting index and defined a developmental curve in the same manner as (i). When a value of the index is larger than 0, children prefer fronting patterns to backing ones. For (iii) and (iv), I defined developmental curves in the same manner as (i), and normalized them.

Moreover, in order to examine the developmental changes in speed of CVCV production, I also calculated mean durations of CVCV sequences for each child for each month of age, and then defined the developmental curves for the mean durations. To clearly distinguish a unit for analysis, I focused only on CVCV rather than both CVC and CVCV.

3.2.2 Analysis of group data for Japanese

To assess developmental trends of group data for Japanese, I pooled data across three children and obtained developmental curves (i)–(iv) (Fig. 3.1) in the same manner as was used for the analysis of individual data. I also obtained a developmental curve of speech rates for

Algorithm 2 The pseudo-code of the kernel regression used in the present study is as follows. In the pseudo-code, a normal (*e.g.*, x), bold symbol (*e.g.*, $\mathbf{x}$), superscript symbol $^{\mathrm{T}}$ and the symbol $\mathbf{I}$ represents a vector, matrix, transpose operator and unit matrix, respectively.

```
INPUT: Independent variables $\mathbf{X} = \{x_i \cdots x_N\}$, predictors $\mathbf{T} = \{t_i \cdots t_N\}$ and the parameter of radial basic function kernel (RBF) $\sigma$, which was determined by leave-one-out cross-validation.
OUTPUT: The predictive variables $\mathbf{Y} = \{y_i \cdots y_N\}$ estimated by this algorithm.
I calculated the Gramian matrix $\mathbf{\Phi}$ using the RBF kernel by following: $\mathbf{\Phi} \leftarrow \{\phi(x_i, \mathbf{X})\}$, where $\phi(a, \mathbf{B}) \leftarrow \exp(-(a - \mathbf{B})^2/\sigma)$.
My aim was to obtain optimal weights coefficient $\mathbf{W}$ of the Gramian matrix $\mathbf{\Phi}$. I assumed the hyper-parameters $\alpha$, which determined variance of the weight coefficient, and $\beta$, which determined variance of noise. In order to obtain optimal $\mathbf{W}$, $\alpha$ and $\beta$, I conducted an iterative optimization. Initially, I set random values to *alpha* and $\beta$, and then conducted following iterative optimization.
repeat
    By using given $\alpha$ and $\beta$, posterior distribution of $\mathbf{W}$ could be calculated by following: $p(\mathbf{W}, \mathbf{T}) = \mathcal{N}(\mathbf{W}|\mathbf{M}, \mathbf{S})$, where the mean $\mathbf{M}$ and variance $\mathbf{S}$ of the gaussian distribution $\mathcal{N}$ could be obtained by $\mathbf{M} \leftarrow \beta\mathbf{S}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}$ and $\mathbf{S}^{-1} \leftarrow \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$.
    I updated $\alpha$ and $\beta$ under given posterior distribution of $\mathbf{W}$ by maximizing marginal likelihood function $p(\mathbf{T}|\alpha, \beta) = \int p(\mathbf{T}|\mathbf{W}, \beta)\, p(\mathbf{W}|\alpha)\, \mathrm{d}\mathbf{W}$. First, I updated $\alpha$ by following equations: $\alpha \leftarrow \gamma/\mathbf{M}^{\mathrm{T}}\mathbf{M}$, and $\gamma \leftarrow \sum \lambda/(\alpha + \lambda)$, where $\lambda$ is eigen vector of $\beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$.
    Second, I updated $\beta$ by following: $\beta \leftarrow \sum(\mathbf{T} - \mathbf{\Phi}\mathbf{M})^2/(N - \gamma)$.
until I conducted this loop 100 times
Finally, I obtained the predicted variables by following: $\mathbf{Y} \leftarrow \mathbf{\Phi}\mathbf{M}$.
```

the group data. In addition to the analysis of children, I examined if ratios of four types of CVC patterns of child-directed speech by mothers change along with those of their children in the group data of Japanese.

3.2.3 Analysis of individual data for English

For the English analysis, I first analyzed the individual data of English in the same manner as was done for the analysis of individual data for Japanese. Since the age of sampling and the number of data vary from child to child, I selected two children whose data cover a wide range of ages and contain a large number of samples (hereafter, I refer them as child 1 and 2, respectively).

3.2.4 Analysis of group data for English

To assess developmental trends of group data for English, I also analyzed pooled data across 21 children in the same manner as was done for the analysis of group data for Japanese.

3.3 RESULTS

3.3.1 Analysis of individual data for Japanese

I conducted individual analysis of two of the three Japanese children. Data of the other one child were not used for the curve fitting because of sparse nature of data sampling. Both children showed early predominance and a later decrease in the case of repetitions (the goodness-of-fit of the kernel regression were R^2=0.765 and 0.573 for child B and C, respectively). Almost all the CVCs were generated by the same place of articulation until around 12 months for child B and 18 months for child C after which the ratio gradually decreased until about 24 months and then it remained the stable (Fig. 3.2).

For the developmental changes in the preference for the direction of articulations (Fig. 3.2ii), although values of the fronting index fluctuated over time, both children showed early preference to the fronting patterns. This preference exists around 12 months and seems to decrease after 18 months (R^2=0.670 and 0.158 for child B and C, respectively). For the development of intra-organ articulations (Fig. 3.2iii), I obtained a developmental curve (R^2=0.833 and 0.278 for child B and C, respectively). For the development of inter-organ articulations (Fig. 3.2iv), I obtained two developmental curves (Fig. 3.2iv), that is, $\mathrm{C_LVC_C}$ + $\mathrm{C_CVC_L}$ (R^2=0.708 and 0.268 for child B and C, respectively) and $\mathrm{C_LVC_D}$ + $\mathrm{C_DVC_L}$ (R^2=0.468 and 0.664 for child B and C, respectively). For child B, normalized ratios of intra-articulatory development were initially higher than those of inter-articulatory development (labial and coronal, and labial and dorsal consonants patterns). Yet, once the ratios of each of labial-coronal and labial-dorsal consonants patterns exceeded those of intra-organ articulations at 19.1 and 14.2 months of age, respectively, they rapidly peaked at 27.6 and 25.1 months of age, respectively. Although actual normalized ratios of child B and C differ from each others, based on patterns of the intersectional and peak months for the intra- and inter-organ articulations, the child C showed a similar pattern of changes, that is, each of the inter-organ articulations exceeded intra-organ ones at 15.1 and 21.0 months of age, respectively, and rapidly peaked at 21.1 and 25.3 months of age, respectively. In addition, I also analyzed the developmental changes in speed of CVCV articulation. Both children showed peaks in the mean duration of CVCV around 18 months after which the value gradually decreased

(Fig. 3.3).

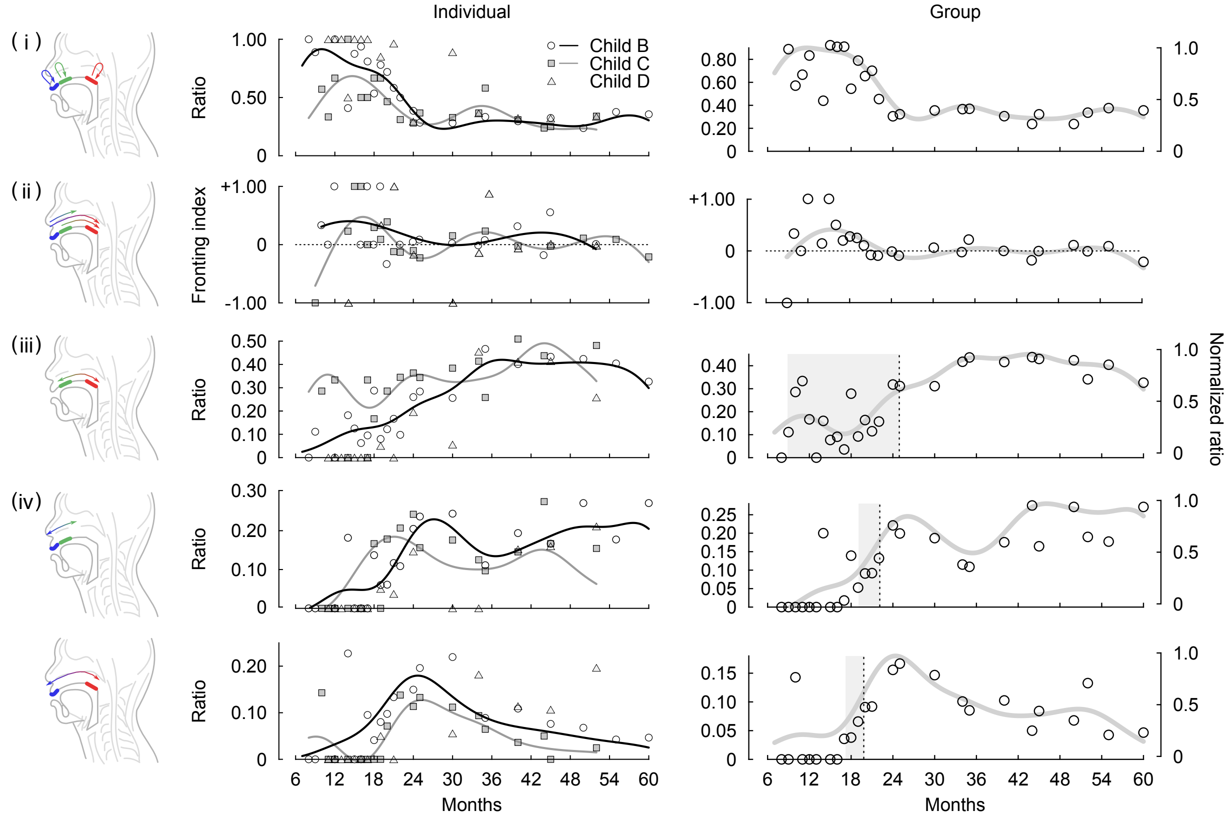

Figure 3.2: The developmental changes in serial order in articulation of consonants obtained by the analysis of individual and group data for Japanese. Middle and right columns show developmental curves obtained by the analysis of individual and group data for Japanese, respectively. Each row indicates (i) repetitions, (ii) fronting, (iii) intra-organ articulations, and (iv) inter-organ articulations. In the middle column, circles, squares and triangles denote relative ratio of each type of CVC patterns produced by child B, child C and child D, respectively. Black, gray and silver curves indicate that developmental curves of child B, child C and child D, respectively. In the right column, circles and lines denotes relative ratio of each type of CVC patterns obtained from pooled data and developmental curves of them, respectively. The shaded areas indicate periods between onset and offset of the developmental changes. I defined the onsets and offsets as months at which a value of curves exceeded 1/3 and 2/3, respectively.

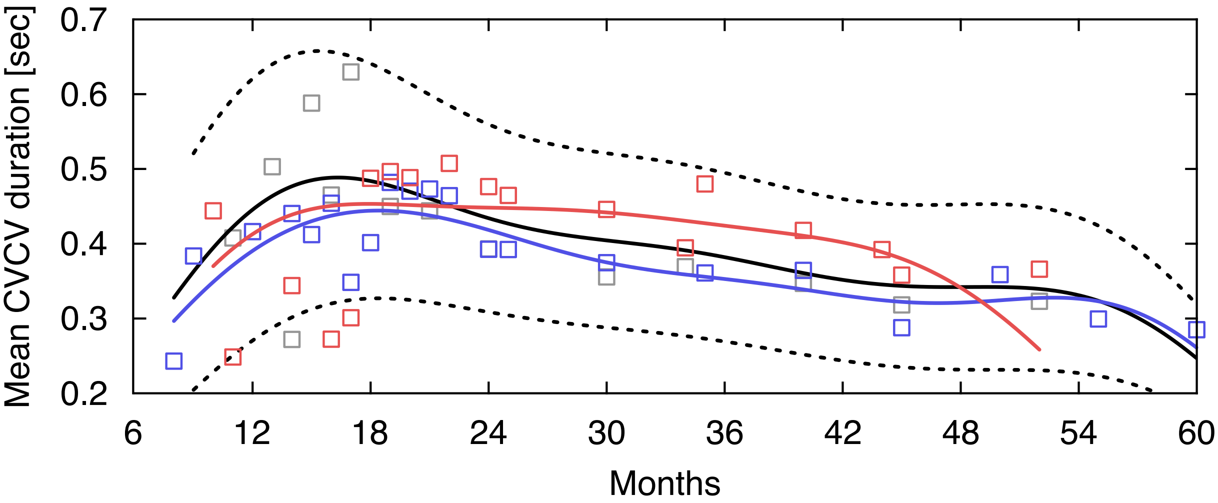

Figure 3.3: The developmental changes of durations of CVCV sequences in Japanese children's speech. The solid and dash lines denote mean durations of CVCV sequences and ±1 standard deviations obtained from the group data. The blue, red and gray markers show the individual analysis for the Japanese children denoting means duration of CVCV sequences.

3.3.2 Analysis of group data for Japanese

In order to assess the developmental trends of group data for Japanese, I analyzed pooled data across three children. For the development of repetitions, I obtained a developmental curve as shown in Fig. 3.2i (R^2=0.722). As the individual analysis has shown, almost all the CVCs were generated at the same place of articulation until around 12 months of age after which the ratio of repetition gradually decreased until 24 months.

For the developmental changes in the preference for the direction of articulations, I obtained a developmental curve of the fronting index as shown in Fig. 3.2ii (R^2=0.200). As a result, I observed that preference to the fronting patterns exists 12 months and peaked around 18 months. This preference gradually decreased over time, and it disappeared after around 24 months. I confirmed that this asymmetry was observed in each of the labial-coronal,

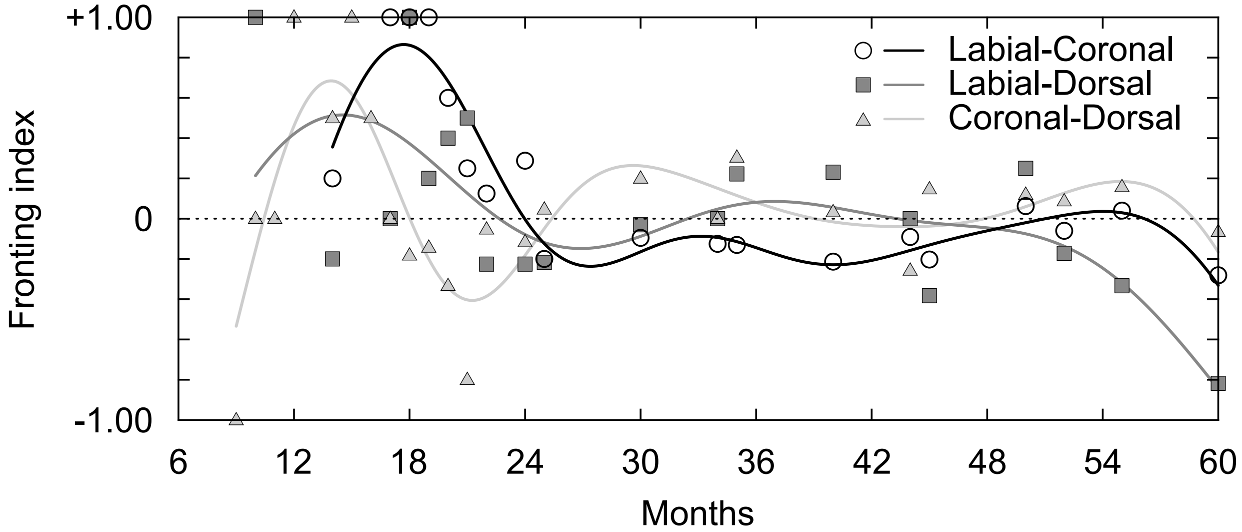

Figure 3.4: The developmental curve of each fronting pattern in Japanese. The black, gray and silver lines show developmental curves of the labial-vowel-coronal, labial-vowel-dorsal, and coronal-vowel-dorsal sequences, respectively. The circles, squares and triangles show raw fronting indices of the labial-vowel-coronal, labial-vowel-dorsal and coronal-vowel-dorsal sequences, respectively.

labial-dorsal and coronal-dorsal consonants patterns (Fig. 3.4). Although the months at which the curves peaked were different among the three types of sequences, the increasing and decreasing patterns were common.

For the development of intra- and inter-organ articulations, I obtained developmental curves as shown in Figs. 3.2iii and iv (R^2=0.752, 0.752, and 0.417 for intra-organ articulations, labial and coronal inter-organ articulations, and labial and coronal inter-organ articulations, respectively). To quantify the developmental trend observed in inter- and intra-organ articulations of Japanese children, I defined durations of development as follows. If a value of the curve exceeded 1/3, I detected this point in time as the onset of develop-

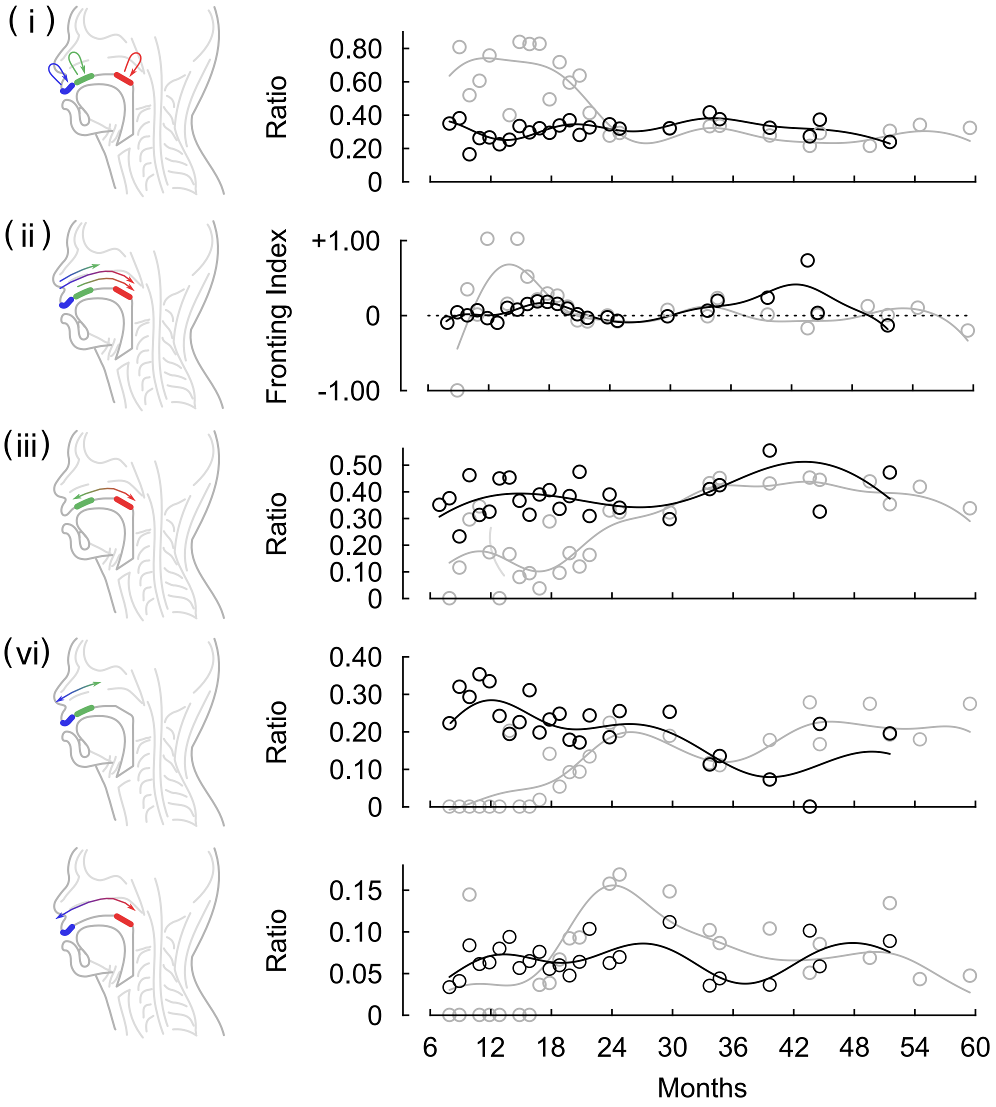

Figure 3.5: The comparison between change in ratios of the repetitions, fronting, intra- and inter-organ articulations in Japanese child-directed and children's speech. Changes in the ratio of occurrences of the repetitions, intra- and inter-organ articulations to the total CVC(V) occurrences, and the fronting index were plotted over months of age for Japanese child-directed speech produced by parents (black) and children's speech (gray).

mental change. The point in time at which the value of curve exceeded 2/3 was detected to determine the offset of the developmental period. The durations of developmental change of intra-organ, labial-coronal inter-organ and labial-dorsal inter-organ were 16.5 (from 8.5 to 25.0), 3.3 (from 18.9 to 22.2) and 3.0 (from 16.9 to 19.9) months, respectively. The analysis showed that the intra-articulator temporal relationship was present early on and before the inter-articulator temporal relationship generated speech production. On the other hand, the developmental change of inter-organ articulations was observed to be faster than that of intra-organ articulations.

Regarding the development of speech rate of Japanese, I found non-linear changes, that is, the mean duration of CVCV increased until around 18 months after which the value gradually decreased (Fig. 3.3). In addition, I found that the developmental changes in repetitions, fronting, intra- and inter-organ articulations of Japanese children's speech were not accompanied by the changes in child-directed speech by their mothers (Fig. 3.5). Yet, the ratios of each type of CVC patterns produced by children tended to converge toward the child-directed speech after 24 months for repetition, fronting, intra-organ articulations and labial-coronal inter-organ articulations and later labial-dorsal inter-organ articulations. Note that, for the child-directed speech, the ratios were globally stable and no preference for the fronting existed.

3.3.3 Analysis of individual data for English

I analyzed the individual data for English and obtained developmental curves of the repetition (R^2=0.470 and 0.822 for child 1 and 2, respectively), the fronting tendencies (R^2=0.0841 and 0.0335 for child 1 and 2, respectively), the intra-organ articulations (R^2=0.310 and 0.517 for child 1 and 2, respectively), the labial-coronal inter-organ articulations (R^2=0.181 and 0.

741 for child 1 and 2, respectively) and the labial-dorsal inter-organ articulations (R^2=0.200 and 0. 311 for child 1 and 2, respectively). I observed that almost all CVCs were generated by the same place of articulation in early development and that the repetitions decreased from 12 to 24 months (Fig. 3.6i). For the developmental changes in the preference for the direction of articulations, although the values of R^2 of both children were very low, raw values of the fronting index showed early preferences to fronting patterns for child 1 and 2, and this preference persisted for child 2 (Fig. 3.6ii). For the development of intra- and inter-organ articulations produced, based on the intersectional months of the normalized ratios of both of articulations, both children showed that normalized ratios of intra-articulatory development were initially higher than those of inter-articulatory development (Figs. 3.6iii and iv). Yet, once the ratios of the labial-coronal and labial-dorsal inter-organ articulations exceeded those of the intra-organ articulations (For child 1, the labial-coronal and labial-dorsal inter-organ articulation exceeded the intra-organ articulations at 12.1 and 11.1 months, respectively. For child 2, they exceeded the intra-organ articulations at 10.7 and 13.2 months, respectively), they rapidly increased.

3.3.4 Analysis of group data for English

I analyzed the group data for English and obtained developmental curves of the repetitions (R^2=0.525), the fronting tendencies (R^2=0.301), the intra-organ articulations (R^2=0.385) and the inter-organ articulations (R^2=0.449 and 0.560 for labial-coronal and labial-dorsal patterns, respectively). The results showed a predominance of repetitions by around 12 months of age and gradually decreased within the range of the tracking periods (Fig. 3.6i).

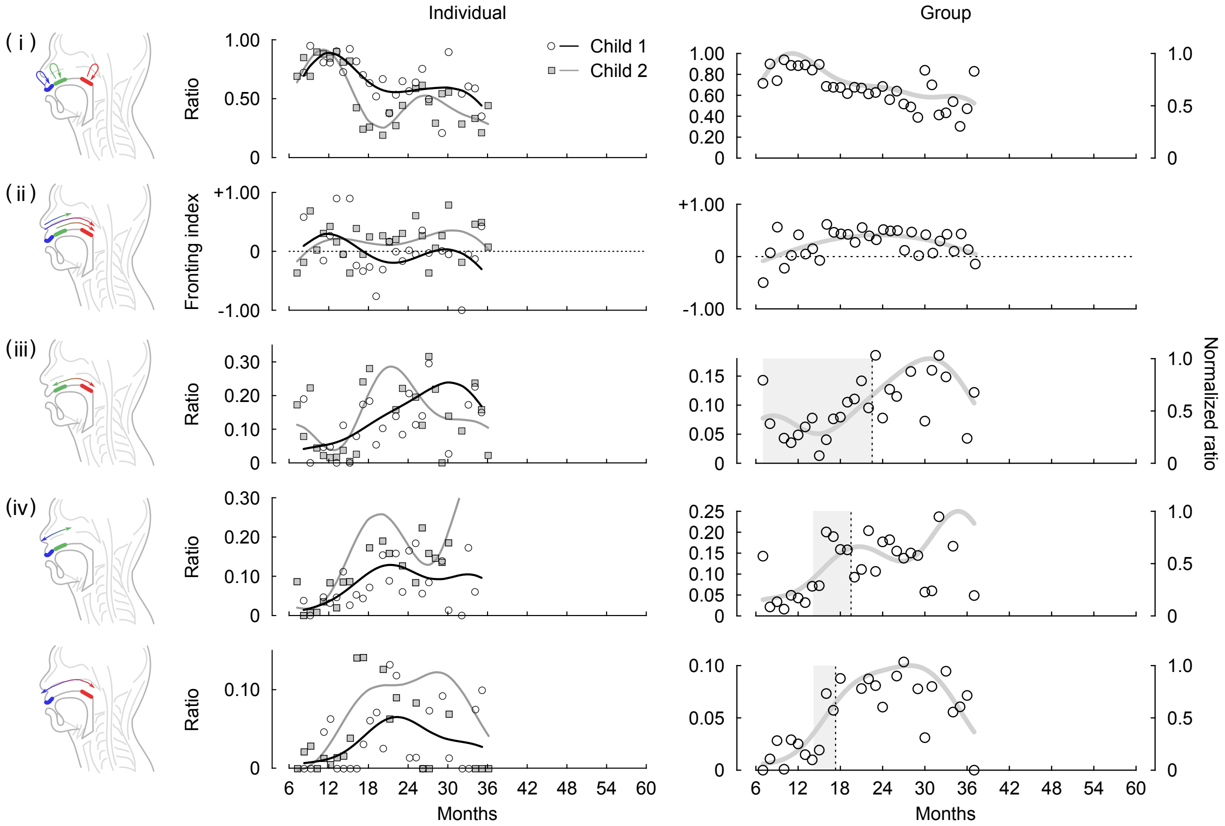

Figure 3.6: The developmental changes in serial order in articulation of consonants obtained by the analysis of individual and group data for English. The Middle and right columns show developmental curves obtained by the analysis of individual and group data for Japanese, respectively. Each row indicates (i) repetitions, (ii) fronting, (iii) intra-organ articulations, and (iv) inter-organ articulations. In the middle column, circles and squares denote relative ratio of each type of CVC patterns produced by child 1 and child 2, respectively. Black and gray curves indicate that developmental curves of child 1 and child 2, respectively. In the right row, circles and curves denotes that relative ratio of each type of CVC patterns obtained by the analysis of group data and developmental curve of them, respectively. The shaded areas indicate periods between onset and offset of the developmental changes. I defined the onsets and offsets as months at which a value of curves exceeded 1/3 and 2/3, respectively.

The preferences to the fronting patterns initially increased and decreased at around 24 months (Fig. 3.6ii). Note that, although previous studies reporting fronting tendencies mainly focus on only labial-coronal patterns [MacNeilage & Davis 2000, MacNeialge 2008], this asymmetry was also observed in labial-dorsal and coronal-dorsal patterns (Fig. 3.7). While the asymmetries of labial-coronal and labial-dorsal pattern persisted until 36 months, those of coronal-dorsal pattern declined to be around zero (Fig. 3.7). As for the intra- and inter-organ articulations, I calculated durations of these developmental changes of them in the same way as was done for the analysis of Japanese. The durations of developmental change of intra-organ, labial-coronal inter-organ and labial-dorsal inter-organ were 15.6 (from

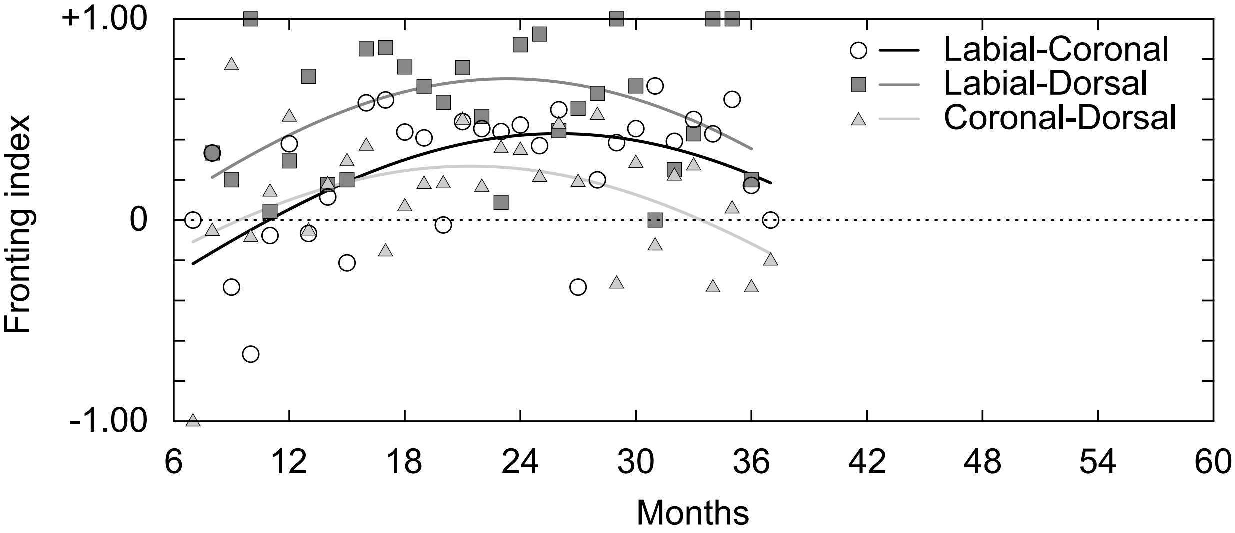

Figure 3.7: The developmental curve of each fronting pattern in English. The black, gray and silver lines show developmental curves of the labial-vowel-coronal, labial-vowel-dorsal, and coronal-vowel-dorsal sequences, respectively. The circles, squares and triangles show raw fronting indices of the labial-vowel-coronal, labial-vowel-dorsal and coronal-vowel-dorsal sequences, respectively.

7.0 to 22.6), 5.9 (from 13.7 to 19.6) and 3.4 (from 14.0 to 17.4) months, respectively. These results also imply later presence and steeper development of the inter-articulatory temporal relationship than the intra-articulatory temporal relationship (Figs. 3.6iii and iv).

3.4 DISCUSSION

In the present study, I analyzed longitudinal data of Japanese children and found common developmental trends in CVC sequences among children. Then, I obtained the developmental curves for the group data of Japanese (Fig. 3.2). The onset and duration of the developmental changes were quantified in each curve for the intra- and inter-organ articulations (Figs. 3.2iii and iv). In the same way as was done for the analysis of Japanese children, I also conducted analysis for English children (Fig. 3.6). As consequences, although actual ratios differ among children and between both languages, I obtained similar developmental trends in both languages. First, until around 18 months, CVC sequences were dominated by repetitions. After 18 months, these sequences decreased. Second, the group analysis showed that, until 24 months, place of articulation was ordered from those produced at the front of the mouth to those produced in the back. Note that English children prefer the fronting patterns after 24 months. Third, CVC sequences generated by movements within the same articulatory organs were already present around 8 months and then gradually increased. Fourth, sequences generated by movements between different articulatory organs tended to appear later but then rapidly increased after the appearance. Finally, I observed a great change around 18 months in not only the ratio of sequences produced by children but also the duration of these sequences.

3.4.1 Repetitive articulation

The results support findings from previous studies [Davis & MacNeilage 1995, Fikkert & Levelt 2008] that infants tend to repeat articulations at the same place of an organ in the period of babbling (Figs. 3.2i and 3.6i). Focusing on the neuromuscular coordination, rhythmic mandibular oscillations have been thought to play an important role in speech production during early development [MacNeilage & Davis 2000, MacNeialge 2008]. Kinematic studies report that infants have independent control only over their jaw and limited control of upper and lower lips especially at one year of age [Green *et al.* 2000, Nip *et al.* 2009]. Earlier maturation of control over the jaw would be consistent with the mandibular oscillation theory, which are produced by the central pattern generators in the brain stem [Gracco & Abbs 1988, Grillner 1982, MacNeialge 2008, Netsell 1982, Wilson *et al.* 2008]. From a phonological perspective, it is unlikely that the infants' linguistic environment cause preferences for repetitions of the same consonants in successive syllables, since preference for variegation of consonants rather than their duplications is generally observed in adult languages [Vihman 1978]. As shown in the Fig. 3.5, this discrepancy between children and adults was also observed in the Japanese child-directed speech contained in the corpus until 24 months. Regarding the preference for variegated sequencing in adult languages, it is argued that variegation would require less energy consumption for the jaw than repetitions: for example, a labial-vowel-coronal sequence (e.g., /pata/ and /tapa/) can be produced by jaw movements with a single cycle, whereas a repetitive sequence (e.g., /papa/ and /tata/) rather requires two jaw cycles [Rochet-Capellan *et al.* 2007]. It can be also argued that, since a discrete movement generating a syllable requires an initiation and termination of movements [Grimme *et al.* 2011], combining two different discrete movements is not always harder than combining the same discrete movements into a sequence. In the present study, I have shown that the ratio of repetitions gradually decreases over time. This result implies that early stability of articulation changes from repetitions to variegations in the developmental process under influence of the linguistic environment.

3.4.2 Preference for direction of articulations

One of the observed characteristics of consonantal changes is the preference for the direction of articulations, namely fronting vs. backing (Figs. 3.2ii, 3.4, 3.6ii and 3.7). Previous studies have reported that children prefer labial-vowel-coronal sequences over coronal-vowel-labial ones [MacNeilage & Davis 2000, MacNeialge 2008]. In the present study, I confirmed that this is indeed the case in children younger than 2 years old for both Japanese and English (Figs. 3.2ii, 3.4, 3.6ii and 3.7). This preferential asymmetry could be attributed to stabilities of phase relationships among movements of the jaw, tongue and lips [Rochet-Capellan *et al.* 2007]. This idea originates from investigations about rhythmic movements of limbs, in which, as a rate of movement increases, the coordination pattern of limbs is destabilized and another stable coordination pattern emerges [Haken *et al.* 1985]. For the articulatory system, as a speech rate increases, phase relationships among movements of articulators cause the preference of a labial-coronal sequence to a coronal-labial one [Rochet-Capellan *et al.* 2007]. This shift is caused by a modification of the coordination between the jaw and constrictors. In these changing processes, labial-to-coronal CVCV disyllables are more favored than coronal-to-labial ones, suggesting that the phase relationship among movements of the jaw, tongue and lips in the former is more stable than the one in the latter. Furthermore, other studies have shown that stop consonantal gestures that occur during fronting of consonant-consonant sequences display more overlap than during backings [Byrd 1996, Chitoran *et al.* 2002, Zsiga 1996].

The preference for fronting becomes lower over development after their peaks. As for Japanese, the preference disappears after 24 months (Figs. 3.2ii and 3.4). This finding is consistent with previous reports that there is no preference for labial-vowel-coronal (LC) in Japanese adult speech production [MacNeilage *et al.* 1999, Tsuji *et al.* 2012]. Together with these studies, the results indicate that Japanese children prefer LC sequences and, over time, they shift their production of these sequences closer to the distribution found in adult speech. A cross-language comparison study of adults found that neuromuscular constraints lead to a universal articulatory bias toward sequences, but that a language-specific perceptual bias emerges from the distributional frequencies found in the native language [Tsuji *et al.* 2012]. Thus, LC sequences have a higher articulatory stability but a lower perceptual stability in Japanese adults. Since infants show LC perceptual bias [Nazzi *et al.* 2009], the developmental change observed in this study implies that speech production is highly constrained by motoric factors in early development and then modulated by the specific language inputs. This adjustment to native languages is also reported in Dutch [Fikkert & Levelt 2008]. Their longitudinal study shows a similar distribution of place of articulation patterns between children's and adult's speech. In fact, although the Japanese preferences of all of the fronting sequences seem to disappear after 24 months, regarding English, the preference to the labial-coronal and labial-dorsal sequences persisted within the range of the tracking periods.

3.4.3 Intra- and inter-articulatory coordination

The main goal of the present study was an examination of the developmental process that shapes the serial order of speech production in terms of intra- and inter-articulatory coordination. A crucial finding of this study is the early emergence of the intra-articulator temporal relationship that produces variegated sequences of consonants in CVC, indicating that changes in the location of the tongue articulation during speech production start in early development (Figs. 3.2iii, 3.2iv, 3.6iii and 3.6iv). Although the mandibular oscillation hypothesis [MacNeilage & Davis 2000, MacNeialge 2008] could explain the early speech production, the finding of the early intra-articulatory coordination requiring independent tongue movements to the jaw suggests that the serial coordination of both the jaw and tongue plays an important role starting in early development [Giulivi *et al.* 2011]. This is further supported by a simulation study showing the role of articulators other than the jaw in a single consonant articulation [Serkhane *et al.* 2007]. Together, these studies imply that the serial coordination of articulators generates a single constrict-release gesture starting in the early developmental stage of speech production. This early presence of the intra-articulator temporal relationship is contrasted with the serial coordination of the inter-articulatory temporal relationships, which emerges later around 18 months for Japanese and 15 months for English. This result implies that serial coordination between the lips and tongue is absent when repetition patterns dominate. Intriguingly, although the onset of the inter-articulator temporal relationship appears later than the intra-articulator one, the developmental change of the inter-articulator relationship is steeper. This observation implies that once children acquire more rapid articulatory movements, stability of the inter-articulator temporal relationship increases.

The analysis of the durations of CVCV sequences revealed that the developmental changes in speech rates are non-linear, that is, the mean duration of CVCV increases until around 18 months, after which the value gradually decreases (Fig. 3.3). In early development, the rhythmic nature of a serial order in speech may be represented as non-segmental articulatory gestures. In other words, the goals of articulatory movements are achieved ambiguously. However, with the development of speech, children generate a clearer serial order by linking together discrete articulatory movements. In the early period of this shift, immature motor control may cause a jerky trajectory. Therefore, the mean duration of CVCV can be comparatively shorter during early development and then gradually increase until around 18 months. Moreover, the development of serial coordination involving inter-articulatory temporal relationships leads to an acceleration of serial order in articulations after around 18 months. Previous studies have reported non-segmental features of speech signals produced by children [Nittrouer 1993] and articulatory rates that increase with age [Walker & Archibald 2006]. These reports are consistent findings in the present study. Changes in the functional domain may be related to anatomical modifications of the vocal tract. In fact, other studies have shown that development of the vocal tract shows a non-linear change around 18 months [Vorperian *et al.* 2005].

3.4.4 Limitations of the present study

In this study, I analyzed individual and group data for both Japanese and English, and obtained developmental curves of repetitions, fronting, intra- and inter-organ articulations

(Figs. 3.2 and 3.6). As a result, I observed some common developmental trends among these articulatory patterns. However, as R^2 values shows particularly in the analysis of individual data, accuracies of some of the curve fitting were not good. The lower goodness-of-fit might be caused by outliers with a small number of samples. Especially, in early development, unreliability of transcription of ambiguous utterances would also lead to lower accuracies of curve fitting. In this sense, direct measurements or estimations of articulatory movements are needed in future studies.

3.5 CONCLUSION

In conclusion, the present study has shown that the development of serial order in speech undergoes great changes around 18 months for Japanese and 15 months for English. Before these periods, a passive synchronization among the jaw-tongue-lip system, mainly driven by the mandibular oscillation, may generate the repetitions. During the same period, the serial coordination of intra-articulator is present but has limited properties. In order to stabilize articulations, serial coordination during this period has a rhythmic nature. Over time, pressure for the child to produce more informative communication induces the rhythmic movements to differentiate into combinations of discrete movements. Furthermore, increased speed of articulations causes the shift in stability, that is, the serial integration of inter-articulators becomes more stable than that intra-articulator serial integration. Thus, after 18 months for Japanese and 15 months for English, articulations produced by different organs emerge and develop rapidly. On the other hand, the serial coordination of intra-articulators exhibit prolonged development to refine the rapid sequences of articulations.

Ultimately, the present study shows the manner by which the developing neuromuscular control of articulatory coordination constrains the serial order in speech production among children in both Japanese and English.

Appendix

In addition to Japanese and English, I also analyzed longitudinal changes in a serial order in Dutch [Fikkert 1994]. The Dutch corpus focuses on only meaningful words, and utterances included in it were recorded under interactions between children and their parents. The calculation of values of the repetition, intra- and inter-organ articulation were same as was done for Japanese and English (see Sec. 3.2). Indices of the fronting tendency were calculated by $(\mathrm{C_LVC_C} + \mathrm{C_LVC_D} + \mathrm{C_CVC_D})/(\mathrm{C_CVC_L} + \mathrm{C_DVC_L} + \mathrm{C_DVC_C})$. Note that, because of a lack of samples at earlier than 12 months in Dutch, onset of the developmental changes in Dutch could not be estimated by the same accuracies as were done for Japanese and English.

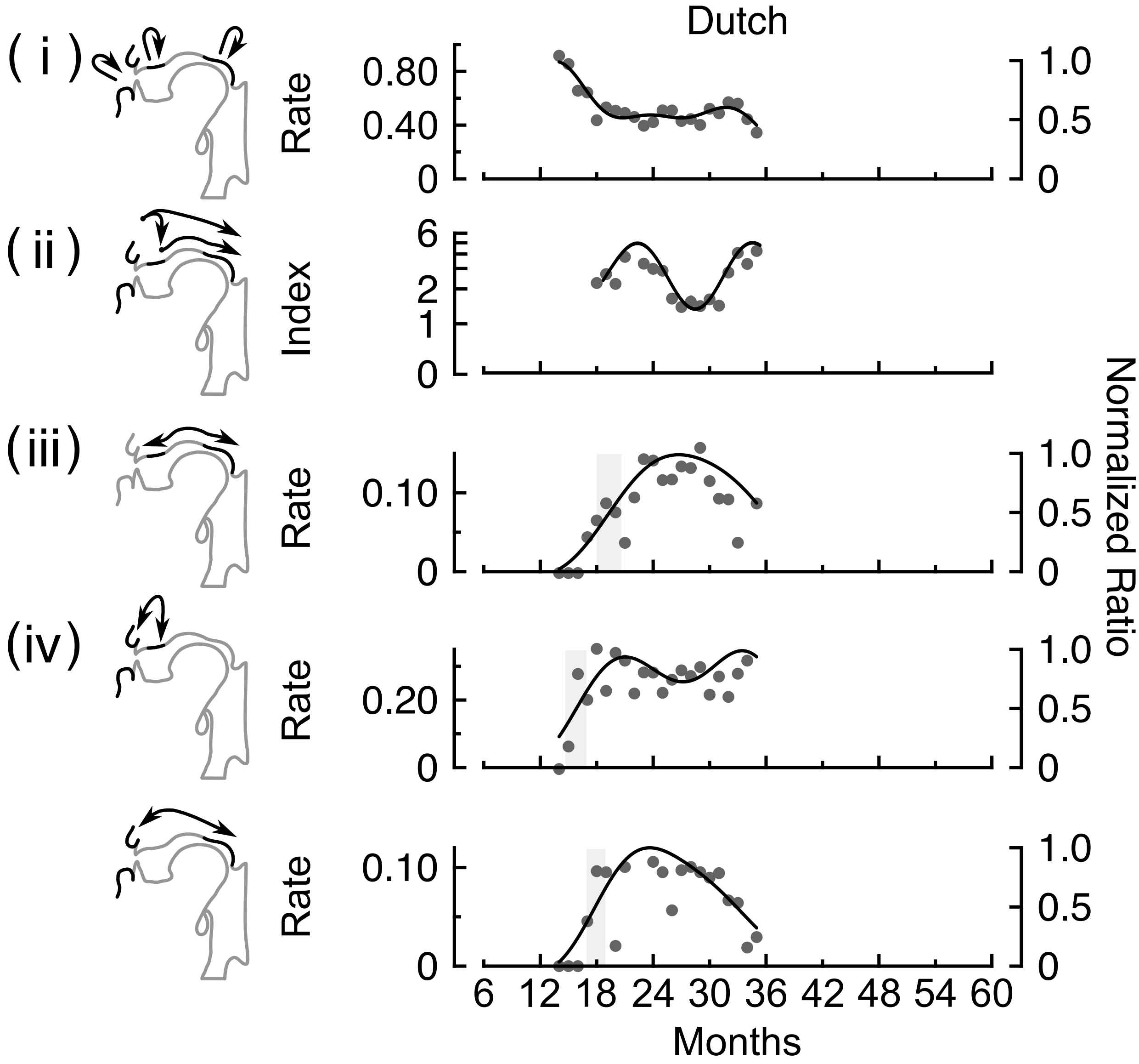

Figure 3.8: The developmental changes in serial order in articulation of consonants obtained by the analysis of group data for Dutch. Each row indicates (i) repetitions, (ii) fronting, (iii) intra-organ articulations, and (iv) inter-organ articulations. In the right column, circles and lines denote relative ratio of each type of CVC patterns obtained from pooled data and developmental curves of them, respectively. The shaded areas indicate periods between onset and offset of the developmental changes. I defined the onsets and offsets as months at which a value of curves exceeded 1/3 and 2/3, respectively.

Chapter 4

General Discussion

4.1 Summary of this thesis

In Chap. 2, I examined the development of vowel articulation using acoustic-to-articulatory inversion. Results of the group analysis showed that development of vowel articulation progressed through the three stages (Fig. 2.5). In the first stage (6–9 months), the articulation of different kinds of vowels depended on a position of the tongue body and lip aperture. In the second stage (10–17 months), articulation depended on the jaw and tongue apex, in addition to the parameters recruited during the first stage. Finally, in the third stage (18 months and older), articulation depended also on the shape of the tongue dorsum. These three stages, observed as part of the group analysis, may have been supported by the individual analysis (Fig. 2.7). In the individual analysis, two shared trends were observed: an early contribution of a position of the tongue body and lip aperture to production of different vowels, and a later contribution of tongue dorsum shape. In the analysis, individual differences were

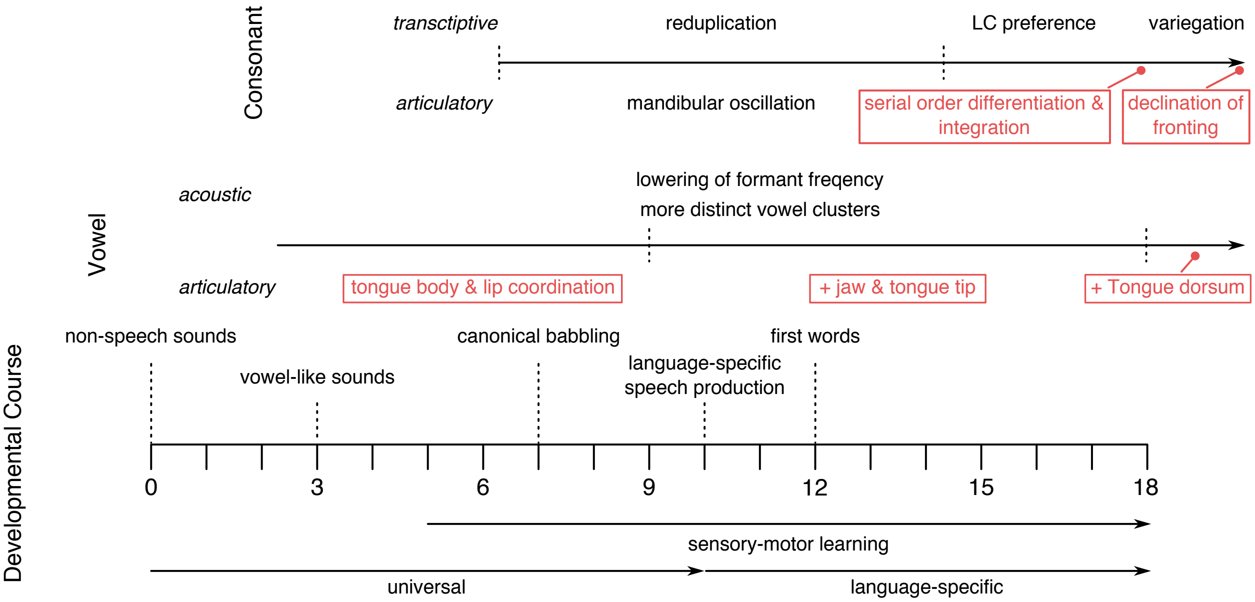

Figure 4.1: Thesis results regarding developmental course.

also observed in the manner of jaw and tongue apex recruitment. Subsequent analyses of the confusion matrix and likelihood of linear discriminant analysis suggested that the three stages reflected differentiation and refinement processes of vowel articulations (Fig. 2.8). In the first stage, vowel articulations were categorized according to three clusters: center, front, and back vowels (see also Fig. 1.8). In the second stage, these clusters became differentiated into five by means of a back-vowel cluster split into /o/ and /u/ clusters, and a front-vowel cluster split into /i/ and /e/ clusters. Eventually, in the third stage, each of these clusters was refined and became more distinct.

In Chap. 3, I examined longitudinal CVC sequence changes in Japanese and English children. Articulatory sequences underlying phonotactics pose a neuromuscular problem; in this sense, relationships between articulatory organs involved in consonant production reflect neuromuscular coordination. Development of serial order in speech underwent significant changes around 18 months for Japanese children, and 15 months for English children (Figs. 3.2 and 3.6). Prior to these periods, a large portion of the CVC sequence consisted of same-consonant repetition, suggesting that passive synchronization among the jaw-tongue-lip system was driven by mandibular oscillation. During these periods, serial intra-articulator coordinations also existed, despite being a minority. Contrary to the intra-articulator coordination, the inter-articulator coordinations appeared in later periods, but their developmental increases were more rapid than those of intra-articulatory coordination. I speculate that these processes reflected the shift in stability induced by the rate of articulations. Moreover, besides implications for developmental changes in articulatory coordination, an alternation caused by external stimuli was also observed. Crosslinguistic differences between Japanese and English were observed in fronting tendencies (Figs. 3.4 and 3.7). Decreases in fronting tendencies were wholly observed in Japanese, but partly observed in English children. Taking into account that fronting tendencies are preferred by coordinations of the articulatory system, a decrease of fronting tendencies would result from external stimuli.

4.2 Development of speech production analyzed in the acoustical space and the articulatory space

Based on physical acoustics, we can infer the geometrical shape of the vocal tract from acoustical features such as formant frequencies. However, because of redundancies of motion in articulatory organs, we cannot always associate acoustical features with a state of articulatory organ. These redundancies prevent us from revealing the developmental course of

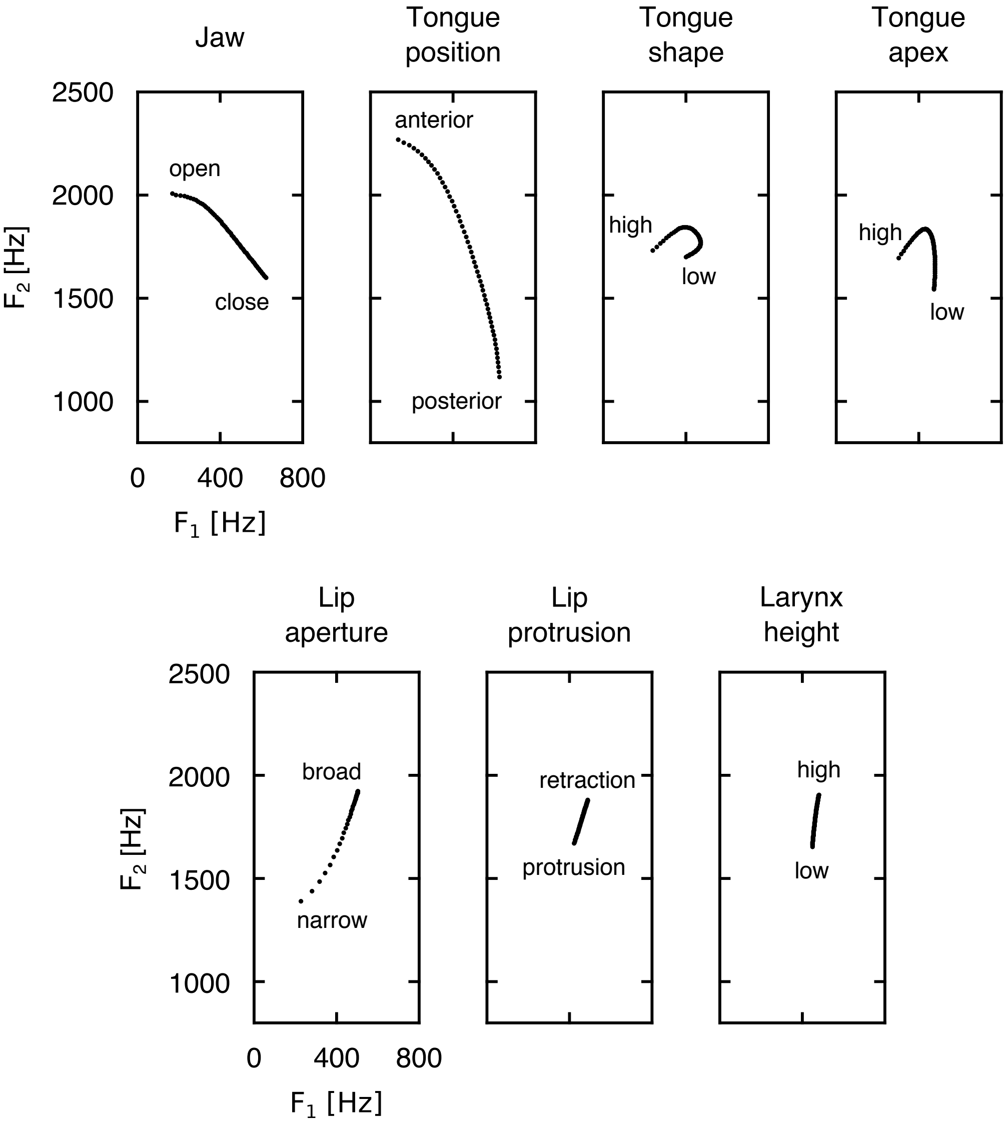

Figure 4.2: A range of formant frequencies generated by a change in each articulatory parameter of the Maeda's model. Formant frequencies were calculated as follows: (i) I set all articulatory parameters to zero; (ii) I changed the value of one of seven articulatory parameters from -3.0 to 3.0 in 0.1 increments. The associations between articulatory parameters with shape of the vocal tract are shown in Fig. 2.1.

speech production. For instance, previous studies [Serkhane *et al.* 2007] have reported that, in the babbling stage, vocalizations produced by infants have more varieties of F_2 than F_1, suggesting that infants explore percepto-motor associations to move the vertical dimension of their vocal tract. However, as shown in Fig. 4.2, many articulatory parameters can generate varieties of F_2 values (i.e., changes in shape of vertical dimension of the vocal tract). With this as background, the next question is which organs infants move during exploration. This question can be answered in the articulatory space. The analysis in Chap. 2 provides part of the answer: a range of a position of the tongue and lip apertures correspond to varieties of vocalizations at the babbling stage at 6–9 months; varieties of vocalizations at the later stage require recruitment of other articulatory movements. This suggests that infants explore percepto-motor associations by coordinating position of the tongue and lip aperture at the early stage, and then recruit other movements to explore the larger acoustic-articulatory mapping space. Note that, as shown by Atal *et al.* (1978) [Atal *et al.* 1978], redundancies in the articulatory space can hamper an accurate estimation of articulatory states. In fact, as shown in Table 2.2, the error of estimation for lip aperture state was large. In this sense, an empirical measurement of articulations is required in future studies.

4.3 Functional and structural development related to the origin of human language

The analysis in Chap. 2 provides some clues as to the origin of human language. The traditional debate on the matter concerns relationships between the development of vocal tract anatomy and motor control. The prevailing view emphasizes anatomical changes in the vocal tract, with laryngeal descent being one origin of human language, in the sense that it enables us to produce quantal vowels [Lieberman 1969, Lieberman 2012]. The other view insists that limitations in infancy are a matter of control rather than anatomy [Boë *et al.* 2013].

It is difficult to discern whether productions of quantal vowels result from laryngeal descent. This occurs because it is hard to define quantal vowels during infancy. In the frequency domain, values of formant frequencies of infant vocalizations are completely different from those in adults. Furthermore, when considering codomain differences in infant and adult formant frequencies, homeomorphism is not necessarily maintained. The articulatory domain lacks research showing the articulatory state of infants during vowel production. Nevertheless, in the present study, I demonstrated that, at least at 6–9 months, infants can generate different articulatory states to produce quantal vowels by coordinating a position of the tongue body and lip aperture.

This finding, alongside a report by Boë *et al* (2013) [Boë *et al.* 2013] (see Sec. 1.2.1), suggests that infants cannot generate sounds acoustically identical to adults' quantal vowels, but can generate articulations homeomorphic to those of adults using quantal vowels. However, it is also doubtful that limitations in infancy are a matter of control rather than anatomy, as Boë *et al.* (2013) concluded. Even if intrinsic muscle fibers of infants' tongues were similar to those of adults, anatomical differences, such as the ratio of vocal tract horizontal length to vertical height, would affect tongue control. In fact, the present study implies the possibility that a broadening of the pharyngeal cavity is involved in the refinement of the vowel cluster.

4.4 Development of serial coordination and underlying neural organizations

In this section, I discuss the development of neural systems underlying longitudinal changes in the serial order revealed in Chap. 3, and consider an approach to solving further issues through the construction of possible models.

Although the tongue is a single physical continuum, it is traditionally divided into three portions: the tongue body, the dorsum, and the apex. Movement characteristics of these parts differ, with muscle fibers varying from part to part. In this sense, the emergence of the intra-organ articulation due to repetition would reflect a differentiation in the neuromuscular coordination of the tongue. On the other hand, inter-organ articulation reflects neuromuscular coordination governing the integration of different organs. As many studies [Gracco & Abbs 1988, Grillner 1982, MacNeialge 2008, Netsell 1982, Wilson *et al.* 2008] claim, simple rhythmic movements such as mandibular oscillations, mastifications, and limb movements are considered to be autonomous behaviors relying on corresponding central pattern generators that work in the brain stem and spinal cord. Contrary to these simple rhythmic movements, complex movements would be based in the cerebral cortex [Bouchard *et al.* 2013, Golfinopoulos *et al.* 2010, Grillner 1982, Price 2010], and dynamic organizations of such movements would contribute to multi-articulator movements.

In spite of recent advances in neuroscience, measuring infants' neural activations during motor productions is a difficult issue, which makes it more challenging to elucidate the dynamics of neural activities that underlie sequential movements and neuromuscular coordi-

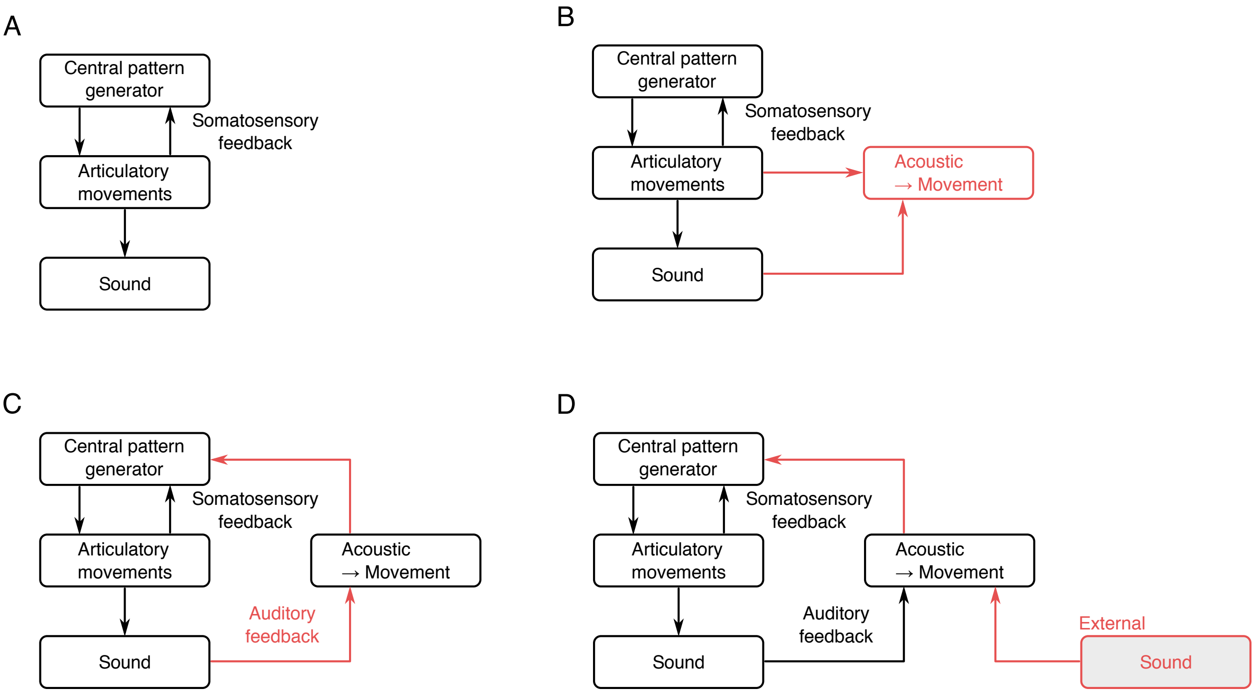

Figure 4.3: Development of speech production hypothesized by this thesis.

nation. One solution could be to construct potential models and explore possible issues [Taga *et al.* 1991]. There have been many attempts to model articulations using an engineering approach to phonetics [Saltzman *et al.* 1989, Sondhi & Schroeter 1987]. Regarding the development of speech production, almost all modeling approaches focus on sensori-motor learning similar to imitation [Guenther 1995]. The same can be said of cognitive science. In this sense, the autonomous motor system underlying mature speech production may have been relatively ignored; a comprehensive model founded on the autonomous motor system is desirable (Fig. 4.3).

Specifically, the autonomous system underlying the repetitive articulation may be modeled as follows. At the kinematic and kinetic level, articulatory organs can be modeled by the finite-element method and mass-spring dynamics [Dang & Honda 2004]; muscle forces can be generated using an extended Hill's model [Hill 1938,Morecki 1987]. Fine-tuned neural dynamics should then be investigated. Repetitive articulations may be thought of as founded on the mandibular oscillation driven by central pattern generators, and passive synergies between the jaw and other articulators (Fig. 4.3A). In addition, it is reported that infant mammals perform many structured limb movements, regardless of wakefulness or sleep. Sensory feedback from these movements is a substantial driver of infant brain activity [Blumberg *et al.* 2013,Petersson *et al.* 2003], suggesting that autonomous movements contribute to motor learning and sensorimotor integration (Fig. 4.3A). At the same time, using auditory-feedback from own vocalizations, acoustic properties may be associated with their corresponding articulations (Fig. 4.3B). Then, the association between acoustic properties and articulations can lead to an alternation of neural activities through auditory-feedback (Fig. 4.3C). A previous study [MacDonald *et al.* 2012] reports that real-time formant perturbation does not alter toddlers' speech, suggesting the possibility that young children do not have real-time auditory-feedback control.

In addition to the development of intrinsic factors, of course, linguistic inputs from infants' environments play an important role in the development of speech production. In accordance with the differentiation and refinement process shown in Chap. 2, and the developmental patterns of fronting articulation showed in Chap. 3, certain universal vocalizations appear to be localized into infants' native language by external inputs. Compared to localizations of phonetic perception to native language, which are achieved after about 6 months of age, differentiation of universal speech production into native language seem to occur later (Fig. 1.15). The results in Chaps. 2 and 3 are consistent with this report. This does not mean, however, that an adaptation of vocalizations begins only after about 10 months, with younger infants having a capacity to adjust their vocalizations to inputs from their environments. I note that it is important to discriminate between imitation and entrainment; imitation meaning children intentionally imitating sounds they hear, and entrainment referring to autonomous oscillatory dynamics, which children have, being incidentally entrained into external inputs. I propose that, once infants acquire the association between acoustic properties and articulations (Fig. 4.3A–C), a neural oscillation established by autonomous sensorimotor systems can be entrained into external sounds through the acquired acoustic-articulatory association (Fig. 4.3D).

4.5 The embodiment

Finally, I discuss the embodiment of speech production development. As I described in Sec. 1.1.1, the classical perspective of generative phonology claims that language independent of its modality [Jackendoff 2002]. The parallel between natural sign language and spoken language [Hickok *et al.* 2001] strongly supports this claim, but there is a possible lack of evolutionary perspectives. Recently, some scholars [Fitch 2013, MacNeialge 2008] have claimed that language originates from motions, such as lip smacking, and visual perception. This would imply that the parallelism does not necessarily support a conception of modality-independent language. As seen in this parallelism, the plasticity is observed in the neural foundation of language. This robust plasticity would be consistent with a solution provided by nature, in that nature does not concern precisely what is used, but harnesses all properties of whatever it employs [Clark 1997]. Thus, regarding production by the articulatory-auditory language system, useful body dynamics should be recruited to simplify operations required for articulation. The central nervous system only processes simplified operations, suggesting that the problem to be solved by the neural system is defined by bodily properties. Moreover, considering language as a medium of communication, properties of both perception and production define the neural problem. In this sense, developmental processes are founded on plasticity, and the dynamics of messy recruitment observed in nature would be the actual principles of development.

Acknowledgement

I am deeply indebted to the many people whose help my student life and to made this thesis possible.

My special thanks go to Professor TAGA Gentaro, whose enormous patients, insightful comments and enthusiastic encouragements were invaluable during my doctoral course. My heartfelt thanks also go to Dr. WATANABE Hama. Without her detailed criticism and encouragement, I could not complete this work.

Professor ICHIKAWA Akira and Dr. SHIROSE Ayako taught me an importance of integrity and enthusiasm to researches. Without their rigid but thoughtful advises, I would quit the college halfway through my master course.

I would like to thank YAGI Kanade, ARITA Ippei, TANAKA Yoshinori, SHONA Chen, SHIN Rei, KITAMURA Taiji, SHONO Yusuke, HOSHI Haruka, HONDA Akiko, HONMA Kentaro, MATSUO Norimichi, MIYAKOSHI Yoji, MOTOJIMA Yoshiro, YAMADA Tetsuji, AMITA Yasuhiro, UEKI Nozomu, ONO Yuji, ONAKA Atsushi, CHIDA Minori, NIO Mariko, FUKUDA Takashi, UCHIDA Daisuke, KUMAGAI Kenzo, SATO Asato, SHIMIZU Ryoma, NAKAYAMA Shintaro, and NOGUCHI Ryo. They are my comrades in the Waseda university and my college life with them is something I'll never forget. Needless to say, encouragements from my colleagues—Dr. GIMA Hirotaka, Dr. FUJII Shinya, Dr. SASAI Shuntaro, KOBAYASHI Yoshio, KANEMARU Nao, IMAI Makiko, KATO Moe and other students of the division of physical and health education— in the university of Tokyo were indispensable for accomplishment of this thesis.

My thanks also go to collaborators and technical assistants at the laboratory I belonged to. ASAKAWA Kayo, KODA Yoshiko and HIRANO Keiko supported my research to undertake a part of my office procedures. Dr. MARUYAMA shin, Professor SAJI Ryoya, Professor NAKANO Hisako and Professor HOMAE Fumitaka gave me warm encouragements at every opportunity.

Professor YAMAMOTO Yoshiharu, Professor NOZAKI Daichi, Professor TOGO Fumiharu, Professor MORITA Kenji, Professor HIRASHIMA Masaya, Professor SASAKI Tsukasa, Dr. NAKAMURA Toru and Dr. HIDAKA Ichiro provided instruction, support, and comments on my studies in many research meetings. I would like to express my sincere thanks to all of them.

Lastly, I would like to thank my family, OOHASHI Masatsugu, Yoko and Maki, and my LEFT hand.

REFERENCES

[**Amano *et al.* 2006**] S. Amano, T. Nakatani, and T. Kondo, "Fundamental frequency of infants' and parents' utterances in longitudinal recordings", J. Acoust. Soc. Am. **119**, 1636–1647 (2006).

[**Amano *et al.* 2009**] S. Amano, T. Kondo, K. Kato, and T. Nakatani, "Development of Japanese infant speech database from longitudinal recordings", Speech Commun. **51**, 510–520 (2009).

[**Atal & Hanauer 1971**] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Am. **50**, 637–655 (1971).

[**Atal *et al.* 1978**] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", J. Acoust. Soc. Am. **63**, 1535–1555 (1978).

[**Bishop 2007**] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York) (2007).

[**Blumberg *et al.* 2013**] M. S. Blumberg, C. M. Coleman, A. I. Gerth, and B. McMurray, "Spatiotemporal Structure of REM Sleep Twitching Reveals Developmental Origins of Motor Synergies", Curr. Biol. **23**, 2100–2109 (2013).

[**Boliek *et al.* 1996**] C. Boliek, T. Hixon, P. Watson, and W. Morgan, "Vocalization and breathing during the first year of life", J. Voice **10**, 1–22 (1996).

[**Bouchard *et al.* 2013**] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation", Nature **495**, 327–332 (2013).

[**Boë *et al.* 2007**] L. J. Boë, J. L. Heim, K. Honda, S. Maeda, P. Badin, and C. Abry, "The vocal tract of newborn human and Neanderthals: Acoustic capacities and consequences for the debate on the origin of language. A reply to Lieberman (2007)", J. Phon. **35**, 564–581 (2007).

[**Boë *et al.* 2013**] L. J. Boë, P. Badin, L. Ménard, G. Captier, B. Davis, P. MacNeialge, T. R. Sawallis, and J. L. Schwartz, "Anatomy and control of the developing human vocal tract: A response to Lieberman", J. Phon. **41**, 379–392 (2013).

[**Browman & Goldstein 1990**] C. P. Browman, and L. M. Goldstein, "Gestural specification using dynamically-defined articulatory structure", J. Phon. **18**, 299–320 (1990).

[**Byrd 1996**] D. Byrd, "Influences on articulatory timing in consonant sequences", J. Phon. **24**, 209–244 (1996).

[**Byrd & Tan 1996**] D. Byrd and C. C. Tan, "Saying consonant clusters quickly", J. Phonetics **24**, 263–282 (1996).

[**Buder *et al.* 2008**] E. H. Buder, L. B. Chorna, D. K. Oller and R. B. Robinson, "Vibratory regime classification of infant phonation", J. Voice **22**, 553–564 (2008).

[**Chiba & Kajiyama 1942**] T. Chiba and M. Kajiyama, *The vowel: Its nature and structure* (Tokyo-Kaseikan, Tokyo) (1942).

[**Chitoran *et al.* 2002**] I. Chitoran, L. M. Goldstein, and D. Byrd, "Gestural overlap and recoverability: Articulatory evidence from Georgian", in *Laboratory Phonology Vol. 7*, edited by C. Cussenhoven, T. Rietfield, N. Warner, (Walter de Gruyter, Berlin) (2002), pp. 419–448.

[**Chomsky & Halle 1968**] N. Chomsky and M. Halle, *The sound pattern of English* (Harper and Row, New York) (1968).

[**Clark 1997**] A. Clark, *Being there: Putting brain, body, and world together again* (MIT Press, Cambridge) (1997).

[**Dang & Honda 2004**] J. Dang and K. Honda, "Construction and control of a physiological articulatory model", J. Acoust. Soc. Am. **115**, 853–870 (2004).

[**Davis & MacNeilage 1995**] B. L. Davis and P. F. MacNeilage, "The articulatory basis of babbling", J. Speech Lang. Hear. Res. **38**, 1199–1211 (1995).

[**Davis 2007**] B. L. Davis, *PhonBank English Davis Corpus* (TalkBank, Pittsburg) (2007).

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

[**de Boysson-Bardies *et al.* 1989**] B. de Boysson-Bardies, P. Halle, L. Sagart, and C. Durand, "A crosslinguistic investigation of vowel formants in babbling", J. Child. Lang. **16** 1–17 (1989).

[**de Boysson-Bardies 1999**] B. de Boysson-Bardies, *How language comes to children: From birth to two years* (MIT Press, Cambridge) (1999).

[**Dehaene-Lambertz *et al.*, 2002**] G. Dehaene-Lambertz, S. Dehaene and L. Hertz-Pannier, "Functional neuroimaging of speech perception in infants", Science **298**, 2013–2015 (2002).

[**Eimas *et al.* 1971**] P. D. Eimas, E. S. Siqueland, P. W. Jusczyk, and J. Vigorito, "Speech perception in infants", Science **171**, 303–306 (1971).

[**Fadiga *et al.* 2002**] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti, "Speech listening specifically modulates the excitability of tongue muscles: A TMS study", Eur. J. Neurosci. **15**, 399–402 (2002).

[**Fant 1960**] G. Fant, *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations* (Mouton, The Hague) (1960).

[**Fikkert 1994**] P. Fikkert, *On the acquisition of prosodic structure* (Holland Academic Graphics, The Hague) (1994).

[**Fikkert & Levelt 2008**] P. Fikkert and C. Levelt, "How does place fall into place? The lexical and emergent constraints in children's developing phonological grammar", in *Contrast in phonology: Theory, Perception, Aquisition*, edited by P. Avery, E. Dresher, and K. Rice, (Mouton de Gruyter, Berlin) (2008), pp. 231–270.

[**Fitch & Giedd 1999**] T. W. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging", J. Acoust. Soc. Am. **106**, 1511–1522 (1999).

[**Fitch 2013**] W. T. Fitch, "Tuned to the rhythm", Nature **494**, 434–435 (2013).

[**Flanagan 1972**] J. L. Flanagan, *Speech analysis, synthesis and perception, 3rd ed.* (Springer-Verlag, New York) (1972).

[**Fujimura 1981**] O. Fujimura, “Temporal organization of articulatory movements as a multidimensional phrasal structure”, Phonetica **38**, 66–83 (1981).

[**Fujisaki 2008**] H. Fujisaki, “In search on models in speech communication research”, in *Proceedings of INTERSPEECH*, (2008).

[**Galantucci *et al.* 2006**] B. Galantucci, C. A. Fowler, and M. T. Turvey, “The motor theory of speech perception reviewed”, Psycon. B. Rev. **13**, 361–377 (2006).

[**Geddes *et al.* 2008**] D. T. Geddes, J. C. Kent, L. R. Mitoulas, and P. E. Hartmann, “Tongue movement and intra-oral vacuum in breastfeeding infants”, Early Hum. Dev. **84**, 471–477 (2008).

[**Gibson *et al.* 2007**] T. Gibson and M. D. Hollywood, “F2 locus equations: Phonetic descriptions of coarticulation in 17- to 22-month-old children”, J. Speech Lang. Haer. Res. **50**, 97–108 (2007).

[**Giulivi *et al.* 2011**] S. Giulivi, D. H. Whalen, L. M. Goldstein, H. Nam, and A. G. Levitt, “An articulatory phonology account of preferred consonant-vowel combinations”, Lang. Learn. Dev. **7**, 202–225 (2011).

[**Goffman & Smith 1999**] L. Goffman and A. Smith, “Development and phonetic differentiation of speech movement patterns”, J. Exp. Psychol. Human **25**, 649–660 (1999).

[**Goldstein 1980**] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children", Ph.D. thesis, Massachusetts Institute of Technology (1980).

[**Golfinopoulos *et al.* 2010**] E. Golfinopoulos, J. A. Tourville, and F. H. Guenther, "The integration of large-scale neural network modeling and functional brain imaging in speech motor control", Neuroimage **52**, 862–874 (2010).

[**Gracco & Abbs 1988**] V. L. Gracco, and J. H. Abbs, "Central patterning of speech movements", Exp. Brain Res. **71**, 515–526 (1988).

[**Green *et al.* 2000**] J. R. Green, C. A. Moore, M. Higashikawa, and R. W. Steeve, "The physiologic development of speech motor control: Lip and jaw coordination", J. Speech Lang. Hear. Res. **43**, 239–255 (2000).

[**Green *et al.* 2002**] J. R. Green, C. A. Moore, and K. J. Reilly, "The sequential development of jaw and lip control for speech", J. Speech Lang. Hear. Res. **45**, 66–79 (2002).

[**Grillner 1982**] S. Grillner, "Possible analogies in the control of innate motor acts and the production of sound in speech", in *Speech motor control*, edited by S. Grillner, B. Lindblom, J. Lubker, and J. Persson, (Pergamon Press, Oxford) (1982), pp. 217–229.

[**Grimme *et al.* 2011**] B. Grimme, S. Fuchs, P. Perrier, and G. Scho oner, "Limb versus speech motor control: A conceptual review", Motor control **15**, 5–33 (2011).

[**Guenther 1995**] F. H. Guenther, “Speech sound ascquisition, coarticulation, and rate effects in a neural network model of speech production”, Psychol. Rev. **102**, 594–621 (1995).

[**Guenther *et al.* 2006**] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, “Neural modeling and imaging of the cortical interactions underlying syllable production”, Brain Lang. **96**, 280–301 (2006).

[**Haken *et al.* 1985**] H. Haken, J. A. S. Kelso, and H. Bunz, “A theoretical model of phase transitions in human hand movements”, Biol. Cybern. **51**, 347–356 (1985).

[**Heinz & Stevens 1965**] J. M. Heinz and K. N. Stevens, “On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech”, in *Proceedings of the 5th International Congress of Acoustics*, (1965), A44.

[**Hickok *et al.* 2001**] G. Hickok, U. Bellugi and E. S. Klima, “Sign language in the brain”, Sci. Am. **284**, 58–65 (2001).

[**Hickok 2012**] G. Hickok, “Computational neuroanatomy of speech production”, Nat. Rev. Neuro. Sci. **13**, 135–145 (2012).

[**Hill 1938**] A. V. Hill, “The heat of shortening and the dynamic constants of muscle”, Proc. R. Soc. London, Ser. B **126**, 136–195 (1938).

[**Hiroya & Honda 2004**] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an HMM-based speech production model”, IEEE Trans. Audio, Speech, Language Process. **12**, 175–185 (2004).

[**Hixon 1971**] T. J. Hixon, "An electromagnetic method for transducing jaw movements during speech", J. Acoust. Soc. Am. **49**, 603–606 (1971).

[**Ho *et al.* 2011**] J. Ho, M. Zañartu, and G. R. Wodicka, "An anatomically based, time-domain acoustic model of the subglottal system for speech production", Speech. Comm. **50**, 1531–1547 (2011).

[**Homae *et al.* 2006**] F. Homae, H. Watanabe, T. Nakano, K. Asakawa, and G. Taga, "The right hemisphere of sleeping infant perceives sentential prosody", NeuroSci. Res. **54**, 276–280 (2006).

[**Honda 1996**] K. Honda, "Organization of tongue articulation for vowels", J. Phonetics **24**, 39–52 (1996).

[**Houde & Jordan 1998**] J. F. Houde, and M. I. Jordan, "Sensorimotor Adaptation in Speech Production", Science **278**, 1213–1216 (1998).

[**Ingram 1974**] D. Ingram, "Fronting in child phonology", J. Child Lang. **1**, 233–241 (1974).

[**Ishizuka *et al.* 2007**] K. Ishizuka, R. Mugitani, H. Kato, and S. Amano, "Longitudinal developmental changes in spectral peaks of vowel produced by Japanese infants", J. Acoust. Soc. Am. **121**, 2272–2282 (2007).

[**Ishizaka & Flaganan 1972**] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell Syst. Tech. J. **51**, 1233–1267 (1972).

[**Itakura 1975**] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. Acoust., Speech, Signal Process. **23**, 67–72 (1975).

[**Jackendoff 2002**] R. Jackendoff, *Foundation of language* (Oxford, New York) (2002).

[**Jakobson 1968**] R. Jakobson, *Child language, aphasia and phonological universals* (Mouton, The Hague) (1968).

[**Jones 2007**] S. S. Jones, "Imitation in infancy: the development of mimicry", Psychol. Sci. **18**, 593–599 (2007).

[**Jusczyk 1997**] P. W. Jusczyk, *Discovery of spoken language* (MIT Press, Cambridge) (1997).

[**Kajikawa *et al.* 2004**] S. Kajikawa, S. Amano, and S. Kondo, "Speech overlap in Japanese mother-child conversation", J. Child. Lang. **31**, 215–230 (2004).

[**Kashino & Kondo 2012**] M. Kashino, and H. M. Kondo, "Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations", Phil. Trans. R. Soc. B **367**, 977–987 (2012).

[**Kelso *et al.* 1984**] J. A. S. Kelso, B. Tuller, E. Vatikiotis-Bateson, and C. A. Flower, "Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures", J. Exp. Psychol. Hum. Percept. Perform. **10**, 812–832 (1984).

[**Kent & Murray 1982**] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months", J. Acoust. Soc. Am. **72**, 353–365 (1982).

[**Kent 2004**] R. D. Kent, "The uniqueness of speech among motor systems", Clin. Phonet. **18**, 495–505 (2004).

[**Kirby *et al.* 2008**] S. Kirby, H. Cornish, and K. Smith, "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language", Proc. Natl. Acad. Sci. **105**, 10681–10686 (2008).

[**Kiritani 1986**] S. Kiritani, "X-ray microbeam method for measurement of articulatory dynamics: Techniques and results", Speech Commun. **5**, 119–140 (1986).

[**Klatt 1980**] D.H. Klatt, "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. **67**, 971–995 (1980).

[**Kondo & Kashino 2007**] H. M. Kondo and M. Kashino, "Neural mechanisms of auditory awareness underlying verbal transformations", Neuroimage **36**, 123–130 (2007).

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

[**Kuhl & Meltzoff 1996**] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: Vocal imitation and developmental change", J. Acoust. Soc. Am. **100**, 2425–2438 (1996).

[**Kuhl 2004**] P. K. Kuhl, "Early language acquisition: craking the speech code", Nature Rev. Neurosci. **5**, 831–843 (2004).

[**Lashley 1951**] K. S. Lashley, "The problem of serial order in behavior", in *Cerebrel mechanism in behavio: The Hixon Symposium*, edited by L. A. Jeffres (Wiley, New York) (1951), pp. 112–131.

[**Lenneberg 1967**] E. H. Lenneberg, *Biological foundations of language* (Wiley, New York) (1967).

[**Liberman 1957**] A. M. Liberman, "Some results of research on speech perception", J. Acoust. Soc. Am. **29**, 117–123 (1957).

[**Liberman *et al.* 1967**] A. M. Liberman, F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy, "Perception of speech code", Psychol. Rev. **74**, 431–461 (1967).

[**Liberman & Mittingly 1985**] A. M. Liberman and I. G. Mittingly, "The motor theory of speech perception revised", Cognition **21**, 1–36 (1985).

[**Liberman & Whalen 2000**] A. M. Liberman and D. H. Whalen, "On the relationship of speech to language", Trends Cogn. Sci. **4**, 187–196 (2000).

[**Lieberman 1969**] P. Lieberman, D. H. Klatt, and W. H. Wilson, "Vocal tract limitations on the vowel repertories", Science **164**, 1185–1187 (1969).

[**Lieberman 2012**] P. Lieberman, "Vocal tract anatomy and the neural bases of talking", J. Phonetics **40**, 608–622 (2012).

[**Liljencrantz & Lindblom 1972**] J. Liljencrants and B. Lindblom, "Numerical simulation of vowel quality systems: The role of perceptual contrast", Language **48**, 839–862 (1972).

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

[**MacDonald *et al.* 2012**] E. N. MacDonald, E. K. Johnson, J. Forsythe, P. Plante and K. G. Munhall, “Children’s development of self-regulation in speech production”, Curr. Biol. **22**, 114–117 (2012).

[**McGurk & MacDonald 1976**] H. McGurk, and J. MacDonald, “Hearing lips and seeing voices”, Nature **264**, 746–748 (1976).

[**MacNeilage *et al.* 1999**] P. F. MacNeilage, B. L. Davis, A. Kinney, and C. L. Matyear, “Origin of serial-output complexity in speech”, Psychol. Sci. **10**, 459–460 (1999).

[**MacNeilage & Davis 2000**] P. F. MacNeilage and B. L. Davis, “On the origin of internal structure of word forms”, Science **288**, 527–531 (2000).

[**MacNeialge 2008**] P. F. MacNeilage, *The origin of speech* (Oxford University Press, Oxford) (2008).

[**MacWhinney 2000**] B. MacWhinney, *The CHILDES project: Tools for analyzing talk, 3rd ed.* (LEA, Mahwah) (2000).

[**Maeda 1982**] S. Maeda, “A digital simulation method of the vocal-tract system”, Speech. Comm. **1**, 199–229 (1982).

[**Maeda 1990**] S. Maeda, “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model”, in *Speech production and speech modeling*, edited by W. L. Hardcastle and A. Marchal (Kluwer Academic, Norwell) (1990), pp. 131–149.

[**Mann & Liberman 1983**] V. A. Mann, and A. M. Liberman, “Some differences between phonetic and auditory modes of perception”, Cognition **14**, 211–235 (1983).

[**Masaki *et al.* 1999**] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Ninomiya, “MRI-based speech production study using a synchronized sampling method”, J. Acoust. Soc. Jpn. (E) **20**, 375–379 (1999).

[**Matyear *et al.* 1998**] C. L. Matyear, P. F. MacNeilage, and B. L. Davis, “Nazalization of vowels in nasal environments in babbling: Evidence for frame dominance”, Phonetica **55**, 1–17 (1998).

[**Mermelstein 1973**] P. Mermelstein, “Articulatory model for the study of speech production”, J. Acoust. Soc. Am. **53**, 1070–1082 (1973).

[**Meltzoff & Moore 1977**] A. N. Meltzoff and M. K. Moore, “Imitation of facial and mannual gestures by human neonates”, Science **198**, 75–78 (1977).

[**Ménard *et al.* 2004**] L. Ménard, J. L. Schwartz, and L. J. Boë, “Role of vocal tract morphology in speech development: Perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood”, J. Speech Lang. Hear. Res. **47**, 1059–1080 (2004).

[**Mokhtari *et al.* 2008**] P. Mokhtari, H. Takemoto, and T. Kitamura, “Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches”, Speech. Comm. **50**, 179–190 (2008).

[**Monnier 2011**] P. Monnier, *Pediatric Airway Surgery* (Springer, New York) (2011).

[**Morecki 1987**] A. Morecki, “Modeling, mechanical description, measurements and control ofthe selected animal and human body manupulation and locomotion movements”, in *Biomechanics of Engineering–Modeling, Simulation, Control*, edited by A. Morecki, (Springer, New York) (1987), pp. 1–28.

[**Nakatani & Irino 2004**] T. Nakatani and T. Irino, “Robust and accurate fundamental frequency estimation based on dominant harmonic components”, J. Acoust. Soc. Am. **116**, 3690–3700 (2004).

[**Nam *et al.* 2009**] H. Nam, L. M. Goldstein, and E. Saltzman, “Self-orgnization of syllable structure: A coupled oscillator model”, in *Approaches to Phonological Complexity*, edited by F. Pellegrino, E. Marisco, and I. Chitoran, (Mouton de Gruyter, Berlin) (2009), pp. 299–328.

[**Nazzi *et al.* 2009**] T. Nazzi, J. Bertoncini, and R. Bijeljac-Babic, “A perceptual equivalent of the labial-coronal effect in the first year of life”, J. Acoust. Soc. Am. **126**, 1440–1446 (2009).

[**Netsell 1982**] R. Netsell, “Speech motor control and selected neurologic disorders”, in *Speech motor control*, edited by S. Grillner, B. Lindblom, J. Lubker, and J. Persson, (Pergamon Press, Oxford) (1982), pp. 247–261.

[**Nip *et al.* 2009**] I. S. B. Nip, J. R. Green, and D. B. Marx, “Early speech motor development: Cognitive and linguistic considerations”, J. Comm. Disorders **42**, 286–298 (2009).

[**Nishimura *et al.* 2003**] T. Nishimura, A. Mikami, J. Suzuki, and T. Matsuzawa, "Descent of the larynx in chimpanzee infants", Proc. Natl. Acad. Sci. **100**, 6930–6933 (2003).

[**Nittrouer 1993**] S. Nittrouer, "The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study", J. Speech Hear. Res. **36**, 959–972 (1993).

[**Nowak & Krakauer 1999**] M. A. Nowak and D. C. Krakauer, "The evolution of language", Proc. Natl. Acad. Sci. **96**, 8028–8033 (1999).

[**Oller 2000**] D. K. Oller, *The emergence of speech capacity* (Lawrence Erlbaum, Mahwah) (2000).

[**Ouni & Laprie 2005**] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion", J. Acoust. Soc. Am. **118**, 444–460 (2005).

[**Pelorson *et al.* 1994**] X. Pelorson, A. HirschBerg, R. R. van Hassel, and A. P. J. Wijnands, "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model", J. Acoust. Soc. Am. **96**, 3416–3431 (1994).

[**Petersson *et al.* 2003**] P. Petersson, A. Waldenström, C. Fåhraeus, and J. Schouenborg, "Spontaneous muscle twitches during sleep guide spinal self-organization", Nature **424**, 72–75 (2003).

[**Petitto *et al.* 2000**] L. A. Petitto, R. J. Zatorre, K. Nikelski, E. J. Dostie, and A. C. Evans, "Speech-like cerebral activity in profoundly deaf people processing signed languages: implication for the neural basis of human language", Proc. Natl. Acad. Sci. **97**, 13961–13966 (2000).

[**Piaget 1969**] J. Piaget and B. Inhelder, *The psychology of the child* (Basic Books, New York) (1969).

[**Price 2010**] C. J. Price, "The anatomy of language: a review of 100 fMRI studies published in 2009", Ann. N. Y. Acad. Sci. **1191**, 62–88 (2010).

[**Prince & Smolensky 2004**] A. Prince, and P. Smolensky, *Optimality Theory: Constrain interaction in generative grammer* (Blackwell, Oxford) (2004).

[**Reilly & Moore 2009**] K. J. Reilly and C. A. Moore, "Respiratory movement patterns during vocalizations at 7 and 11 months of age", J. Speech Lang. Hear. Res. **52**, 223–239 (2009).

[**Rochet-Capellan *et al.* 2007**] A. Rochet-Capellan and J. L. Schwartz, "An articulatory basis for the labial-to-coronal effect: /pata/ seems a more stable articulatory pattern than /tapa/", J. Acoust. Soc. Am. **121**, 3740–3754 (2007).

[**Sahin *et al.* 2009**] N. T. Sahin, S. Pinker, S. S. Cash, D. Schomer, and E. Halgren, "Sequential processing of lexical, grammatical, and phonological information within Broca's area", Science **16**, 445–449 (2009).

[**Saltzman *et al.* 1989**] E. L. Saltzman, and K. G. Munhall, "A dynamical approach to gestural patterning in speech production", Haskins Laboratories Status Report on Speech Research **100**, 38–68 (1989).

[**Sasaki *et al.* 1977**] C. T. Sasaki, P. A. Levine, J. T. Laitman, and E. S. Jr. Crelin, “Postnatal descent of the epiglottis in man. A preliminary report”, Arch. Otolaryngol. **103**, 169–171 (1977).

[**Sato *et al.* 2006**] M. Sato, J. L. Schwartz, C. Abry, M. A. Cathiard, and H. Lœvenbruck, “Multistable syllables as enacted percepts: A source of an asymmetric bias in the verbal transformation effect”, Percept. Psychophys. **68**, 458–474 (2006).

[**Sato *et al.* 2009**] M. Sato, P. Tremblay, and V. L. Gracco, “A mediating role of the premotor cortex in phoneme segmentation”, Brain Lang. **111**, 1–7 (2009).

[**Schünke *et al.* 2007**] M. Schünke, E. Schulte, U. Schumacher, M. Voll, and K. Wesker, *Thieme Atlas of Anatomy: Head and Neuroanatomy* (Theime, New York) (2007).

[**Schwartz *et al.* 1997**] J. L. Schwartz, L. J. Boë, N. Vallée, and C. Abry, “The Dispersion-Focalization theory of vowel systems”, J. Phon. **25**, 255–286 (1997).

[**Serkhane *et al.* 2007**] J. E. Serkhane, J. L. Schwartz, L. J. Boë, B. L. Davis, and C. L. Matyear, “Infants’ vocalizations analyzed with an articulatory model: A preliminary report”, J. Phonetics **35**, 321–340 (2007).

[**Shirai 1993**] K. Shirai, “Estimation and generation of articulatory motion using neural networks”, Speech Commun. **13**, 45–51 (1993).

[**Smith & Zelaznik 2004**] A. Smith and H. N. Zelaznik, “Development of functional synergies for speech motor coordination in childhood and adlescence”, Dev. Psychol. **45**, 1077–1087 (2004).

[**Sondhi & Schroeter 1987**] M. M. Sondhi and J. S. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer", IEEE Trans. Acoust. Speech Sig. Proc. **35**, 955–967 (1987).

[**Soquet *et al.* 2002**] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI", Speech Commun. **36**, 169–180 (2002).

[**Stoel-Gammon & Cooper 1984**] C. Stoel-Gammon and J. A. Cooper, "Patterns of early lexical and phonological development", J. Child Lang. **11**, 247–271 (1984).

[**Story 2002**] B. H. Story, "An overview of the physiology, physics and modeling of the sound source for vowels", Acoust. Sci. & Tech. **23**, 195–206 (2002).

[**Story 2009**] B. H. Story, "Vowel and consonant contributions to vocal tract shape", J. Acoust. Soc. Am. **126**, 3231–3254 (2009).

[**Sussman *et al.* 1996**] H. M. Sussman, F. D. Minifie, E H. Buder, and C. Stoel-Gammon, "Consonant-vowel interdependencies in babbling and early words", J. Speech Lang. Haer. Res. **39**, 424–433 (1996).

[**Taga *et al.* 1991**] G. Taga, Y. Yamaguchi, and H. Shimizu, "Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment", Biol. Cybern. **65**, 147–159 (1991).

[**Takemoto 2001**] H. Takemoto, "Morphological analyses of the human tongue musculature for three-dimensional modeling", J. Speech Lang. Hear. Res. **44**, 95–107 (2001).

[**Takano & Honda 2007**] S. Takano and K. Honda, “An MRI analysis of the extrinsic tongue muscles during vowel production”, Speech Commun. **49**, 49–58 (2007).

[**Takemoto 2008**] H. Takemoto, “Morphological analyses and 3D modeling of the tongue musculature of the chimpanzee (Pan troglodytes)”, Am. J. Primatol. **70**, 966–975 (2008).

[**Titze & Story 2002**] I. R. Titze and B. H. Story, “Rules for controlling low-dimensional vocal fold models with muscle activation”, J. Acoust. Soc. Am. **112**, 1064–1076 (2002).

[**Trembley *et al.* 2003**] S. Trembley, D. M. Shiller, and D. J. Ostry, “Somatosensory basis of speech production”, Nature **423**, 866–869 (2003).

[**Tsuji *et al.* 2012**] S. Tsuji, N. Gonzalez-Gomez, V. Medina, T. Nazzi, and R. Mazuka, “The labial-coronal effect revised: Japanese adults say pata, but hear tapa”, Cognition **125**, 413–428 (2012).

[**Vallabha & Tuller 2002**] G. K. Vallabha and B. Tuller, “Systematic errors in the formant analysis of steady-state vowels”, Speech Commun. **38**, 141–160 (2002).

[**Vihman 1978**] M. M. Vihman, “Consonant harmony: Its scope and function in child language”, in *Universal Human Language: Vol. 2. Phonology*, edited by J. H. Greenberg, (Stanford University Press, Stanford) (2008), pp. 281–334.

[**Vilain *et al.* 1999**] A. Vilain, C. Abry, P. Badin, and S. Brosda, “From idiosyncratic pure frame to variegated babbling: Evidence for articulatory modeling”, in *Proceedings of the 14th International Congress of Phonetic Science* (1999), pp. 2497–2550.

[**Vorperian *et al.* 1999**] H. K. Vorperian, R. D. Kent, L. R. Gentry, and B. S. Yandell, "Magnetic resonance imaging procedures to study the concurrent anatomic develoment of vocal tract structure: preliminary results", Int. J. Pediatr. Otorhi **49**, 197–206 (1999).

[**Vorperian *et al.* 2005**] H. K. Vorperian, R. D. Kent, M. J. Lindstörm, C. M. Kalina, L. R. Gentry, and B. S. Yandell, "Development of vocal tract length during early childhood: A magnetic resonance imaging study", J. Acoust. Soc. Am. **117**, 338–350 (2005).

[**Vorperian & Kent 2007**] H. K. Vorperian and R. D. Kent, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data", J. Speech Lang. Hear Res. **50**, 1510–1545 (2007).

[**Wakita 1973**] H. Wakita, "Direct estimation of the vocal tract shape by inversion filtering of acoustic speech waveforms", IEEE Trans. Audio Electroacoust. **21**, 417–427 (1973).

[**Walker & Archibald 2006**] J. F. Walker and L. M. Archibald, "Articulation rate in preschool children: A 3-year longitudinal study", Int. J. Lang. Commun. Disord. **41**, 541–565 (2006).

[**Whalen *et al.* 2007**] D. H. Whalen, A. G. Levitt and L. M. Goldstein, "VOT in the babbling of French- and English-learning infants", J. Phon. **35**, 341–352 (2007).

[**Wilson *et al.* 2008**] E. M. Wilson, J. R. Green, Y. Y. Yunusova, and C. A. Moore, "Task specificity in early oral motor development", Semin. Speech Lang. **29**, 257–266 (2008).

[**Yates 1963**] A. J. Yates, "Delayed auditory feedback", Psychol. Bull. **60**, 213–232 (1963).

[**Zharkova *et al.* 2011**] N. Zharkova, N. Hewlett, and W. J. Hardcastle, "Coarticulation as an indicator of speech motor control development in children: An ultrasound study", Motor Control **15**, 118–140 (2011).

[**Zsiga 1996**] E. C. Zsiga, "Acoustic evidence for gestural overlap in consonant sequence", J. Phon. **22**, 121–140 (1996).

[**Zuidema & deBoer 2009**] W. Zuidema and B. de Boer, "The evolution of combinational phonology", J. Phon. **37**, 125–144 (2009).

[**http://www.bartleby.com/107/**] Gray, Henry. 1918. Anatomy of the Human Body, `http://www.bartleby.com/107/`.

[**http://www.cns.bu.edu/ speech/VTCalcs.php**] The Speech Lab / Department of Cognitive and Neural Systems / Boston University, `http://www.cns.bu.edu/~speech/VTCalcs.php`, [Online; accessed 30-Dec-2012].